

I've Fallen and I Can't Get Up: Can High-Ability Students Recover From Early Mistakes in CAT?

By: Kelly L. Rulison and Eric Loken

[Rulison, K. L.](#) & Loken, E. (2009). I've fallen and I can't get up: Can high ability students recover from early mistakes in Computer Adaptive Testing? *Applied Psychological Measurement*, 33, 83-101. DOI: 10.1177/0146621608324023

Made available courtesy of Sage Publishing: <http://apm.sagepub.com/>

*****Reprinted with permission. No further reproduction is authorized without written permission from the Sage. This version of the document is not the version of record. Figures and/or pictures may be missing from this format of the document.*****

*****Note: This version of the document is not the copy of record.**

Abstract:

A difficult result to interpret in Computerized Adaptive Tests (CATs) occurs when an ability estimate initially drops and then ascends continuously until the test ends, suggesting that the true ability may be higher than implied by the final estimate. This study explains why this asymmetry occurs and shows that early mistakes by high ability students can lead to considerable underestimation, even in tests with 45 items. The opposite response pattern, where low-ability students start with lucky guesses, leads to much less bias. The authors show that using Barton and Lord's four-parameter model (4PM) and a less Lord's four-parameter model (4PM) and a less informative prior can lower bias and root mean square error (RMSE) for high-ability students with a poor start, as the CAT algorithm ascends more quickly after initial underperformance. Results also show that the 4PM slightly outperforms a CAT in which less discriminating items are initially used. The practical implications and relevance for psychological measurement more generally are discussed.

Index terms: computerized adaptive testing, Bayesian, item response theory, achievement testing, high-stakes assessment

Computerized adaptive tests (CATs) combine item response theory (IRT) models with real time estimation algorithms to provide tailored assessments that can improve measurement efficiency and reduce examinee burden (Chang, 2004; Meijer & Nering, 1999; Wainer et al., 2000; Weiss, 1982). Compared with nonadaptive "paper and pencil" measures, CATs are more efficient because they update the latent trait estimate, $\hat{\theta}$ after each response and then adaptively select the most appropriate item to deliver next. Obtaining unbiased and efficient estimates of θ in a CAT requires (a) an underlying IRT model that closely corresponds to respondent behavior (van Krimpen-Stoop & Meijer, 2000; Wainer et al., 2000) and (b) an effective item selection algorithm (Chang & Ying, 1996, 1999; Passos, Berger, & Tan, 2007). (Chang & Ying, 1996, 1999; Passos, Berger, & Tan, 2007).

The authors focus on the first requirement as it relates to a specific estimation problem in CATs. Ideally, $\hat{\theta}$ should reach the neighborhood of the true θ before the CAT concludes. In a typical test, ability estimates for a high-ability student might initially ascend quickly and then oscillate a little above and below the final $\hat{\theta}$ as the student encounters questions that are closely matched to his or her ability. In some cases, however, $\hat{\theta}$ may drop at the beginning of the test, and then ascend continuously until the final estimate, suggesting that the student's true θ is perhaps significantly higher than the final $\hat{\theta}$.

In the first simulation study, it will be demonstrated that a pattern of continuously ascending ability estimates can arise when a high-ability student misses early items in a CAT. Under the widely used three-parameter model (3PM), the lower asymptote is a nonzero value that accounts for the possibility of guessing the correct answer on

multiple choice tests. However, the upper asymptote for the item response function is 1, suggesting that a high-ability student should answer an easy question with probability approaching 1. It is conceivable, however, that $P(\theta) = 1$ may not always hold, even if the item appears too easy for the respondent. High-ability students who are anxious, distracted by poor testing conditions, unfamiliar with computers, careless, or who misread the question, may on occasion miss items that they otherwise should have answered correctly. If this happens early in the test, it may lead to the problematic outcome in which the estimates are increasing even at the end of the test.

The potential for underestimation of high-ability students in CAT is rarely discussed in the literature. Most research on obtaining unbiased estimates of θ focuses on identifying aberrant response patterns through person misfit indices (van Krimpen-Stoop & Meijer, 2000) or on adapting item selection algorithms to reflect the uncertainty that exists as the test begins (Chang & Ying, 1996, 1999, 2002; Passos et al., 2007). Chang and Ying (2002) and Chang (2004), for example, argue that item selection algorithms based solely on Fisher's information criterion select items with high a parameters first, yielding step sizes for $\hat{\theta}$ that are inappropriately large at the onset of a CAT. They suggest using an item selection strategy that stratifies the item pool and uses less discriminating items early in the test. This stratification ensures that enough high-discriminating items are left in the item pool to allow $\hat{\theta}$ to ascend quickly at the end of the test (Chang & Ying, 1999, 2002).

Next, another approach is considered. The authors hypothesize that making minor adjustments to the commonly used 3PM may protect against underestimation when a student starts the test to the commonly used 3PM may protect against underestimation when a student starts the test poorly. Specifically, a 4PM proposed by Barton and Lord (1981) is revisited and the authors argue that it might have some utility in CAT.

The Three-Parameter Model (3PM)

In IRT, the probability of a correct response is modeled as a function of a latent trait, θ , and item parameters. The 3PM is frequently used in academic testing and is given by

$$P_j(\theta) = P(X_j = 1 | \theta; a_j, b_j, c_j) = c_j + (1 - c_j) \frac{e^{1.7a_j(\theta - b_j)}}{1 + e^{1.7a_j(\theta - b_j)}}, \quad (1)$$

where a_j is the item discrimination or "slope" parameter, b_j is the item threshold or "difficulty" parameter, and c_j is the lower asymptote or "pseudo-guessing" parameter. The nonzero lower asymptote admits the possibility that low-ability students may occasionally guess the correct answer to difficult items. In contrast, the upper asymptote of 1 reflects the stiff assumption that if an item is easy enough relative to a student's ability, then the probability of a correct response is effectively 1.

Assuming the item parameters are known, the likelihood function for a response vector \mathbf{x} , indicating correct and incorrect responses is given by

$$P(\mathbf{x} | \theta; \mathbf{a}, \mathbf{b}, \mathbf{c}) = \prod_{j=1}^n P_j(\theta)^{x_j} (1 - P_j(\theta))^{1 - x_j}. \quad (2)$$

Choosing $\hat{\theta}$ to maximize equation (2) yields the maximum likelihood estimator (MLE).

Alternatively, in Bayesian estimation, multiplying the likelihood function by a prior distribution yields the posterior distribution:

$$p(\theta | \mathbf{x}) \propto p(\theta) P(\mathbf{x} | \theta; \mathbf{a}, \mathbf{b}, \mathbf{c}): \quad (3)$$

Bayesian estimation in IRT is widely used (Baker & Kim, 2004; Bock & Mislevy, 1982), most often with $p(\theta) \sim N(0, 1)$ and taking either the posterior mean (expected a posteriori or EAP) or posterior mode (modal a posteriori or MAP) to estimate $\hat{\theta}$. The benefits of the Bayesian approach include guaranteed proper estimates

and smaller standard errors, at the cost of some bias in the tails of the ability distribution because of pull from the prior (Baker & Kim, 2004).

Difficulties in Estimating Ability

Some response patterns, such as getting all items correct or incorrect, yield improper estimates for the MLE. In CATs, ad hoc measures are required to implement a floor or ceiling for $\hat{\theta}$ in the early stages of the test. But even if the student has made both correct and incorrect responses, some response patterns can still yield improper estimates. It is worth exploring these patterns because they are related to the underestimation problem when high-ability students miss easy questions early in a CAT.

Consider responses to a three-item test, two correct and one incorrect. The likelihood is

$$p(\mathbf{x} = (1, 1, 0) | \theta) = P_1(\theta)P_2(\theta)(1 - P_3(\theta)) = P_1(\theta)P_2(\theta) - P_1(\theta)P_2(\theta)P_3(\theta). \quad (4)$$

If all items have $a_j = 1.1$ and $c_j = 0.2$, the likelihood is bounded for low y at $c^2 - c^3$ and goes to 0 for high θ . What happens in between depends on the relative difficulty of the items.

Figure 1 shows the likelihood when $b_1 = -1$, $b_2 = 0$, and $b_3 = 1$. If the student answers the easy and moderate items correctly and misses the hardest item (solid line), the MLE is $\hat{\theta} = 0.46$. If, however, the easiest item is missed and the moderate and hard items are answered correctly (i.e., if $\mathbf{x} = (0, 1, 1)$), the MLE is improper, tailing off to negative infinity (dashed line). The likelihood decreases monotonically because the term $(1 - P_1(\theta))$ begins to drop to 0 faster than $P_2(\theta)P_3(\theta)$ increases from the lower asymptote. As long as the incorrect item is significantly easier than the other two items, the rise to $P_1(\theta) = 1$ (and thus the decrease of $1 - P_1(\theta)$ to 0) dominates the likelihood function.

The result is also explained by Bradlow's (1996) demonstration that in the 3PM, the observed information for an item response can actually be negative. Although the expected information for an item is always positive, the observed information provided by a correct response can be negative under certain conditions. Bradlow showed that negative information will occur when an item is answered correctly and

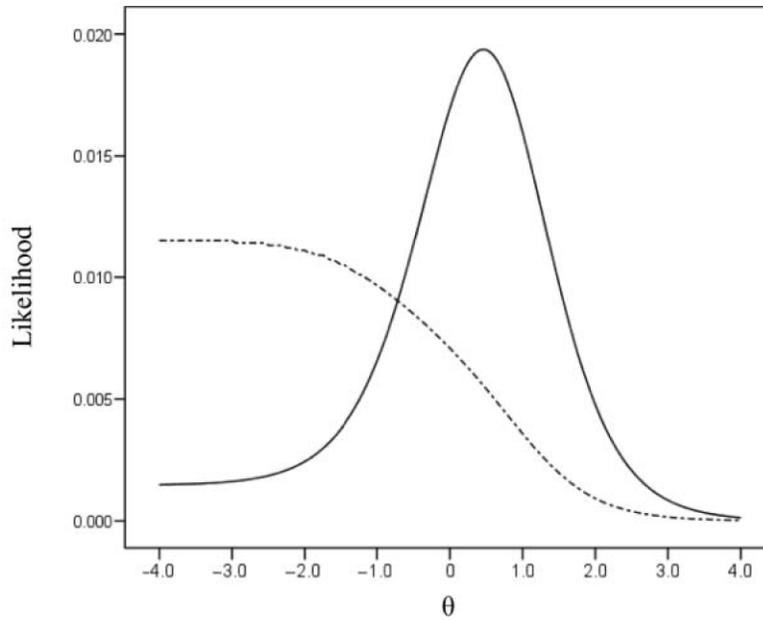
$$\theta < b + \frac{1}{2a} \log c. \quad (5)$$

The likelihood shown by the dashed line is not bounded for low θ , because the information provided by the correct answers to questions 2 and 3 is negative. Therefore, for low y , the observed responses do not give adequate information for a proper ability estimate.

With a proper prior, the EAP for the three-item example is finite, regardless of the pattern of responses. If $p(\theta) \sim N(0, 1)$ and the same response pattern that yielded the improper likelihood in Figure 1 (dashed line) is observed, $\hat{\theta} = -0.4$, $SD = 0.94$ (Figure 2, solid line). But even though the Bayesian estimates are proper, abnormal response patterns can still yield surprising results. For example, if the difficulty of the correctly answered items were raised to $b_2 = b_3 = 2.5$ (with the incorrect item still at $b_1 = -1$), it might be expected that $\hat{\theta}$ would increase. Instead, the posterior mean shifts lower to $\hat{\theta} = -0.91$ and the standard deviation shrinks to 0.91 (Figure 2, dashed line).

That $\hat{\theta}$ shifts lower, with greater certainty, is at first counterintuitive, as correctly answering very difficult questions would seem to indicate high ability. However, because $1 - P_1(\theta)$ goes to 0 much faster than $P_2(\theta)$ and $P_3(\theta)$ rise from c , the information gathered from the more difficult items is

Figure 1
Proper and Improper Likelihood Functions for Three-Item Test



Note. Likelihood for a three-item test, in which each item has $a_j = 1.1$, $c_j = 0.2$, and $b_j = -1, 0,$ and $+1$. The solid line shows the likelihood when only the easiest item is answered incorrectly. The dashed line shows the likelihood when only the hardest item is answered incorrectly.

discounted; informally one might say that the correct answers to highly difficult items “confirm” that they were “just guesses.” In terms of Bradlow’s (1996) discussion, the observed information becomes even more negative if the correct items are more difficult relative to ability, and thus the increase in the posterior variance.

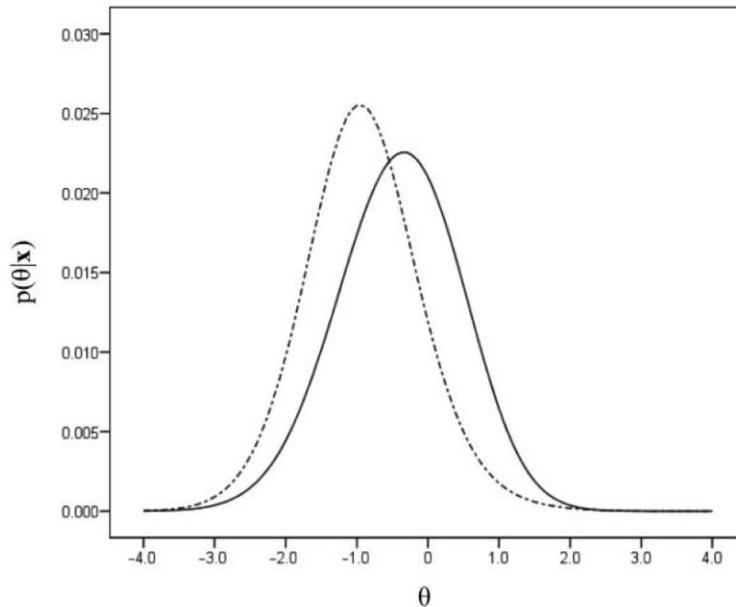
The problematic combination of incorrect answers to easy items with correct answers to harder items has received some attention in the traditional nonadaptive testing literature. Mislevy and Bock (1982) explored the use of ability estimators that were robust to guessing by low-ability students and “carelessness” by high-ability students. They proposed down-weighting responses to items that appeared to be too easy or difficult for the student, given the final $\hat{\theta}$, and completely trimming items that were far from the student’s final $\hat{\theta}$. The weighting procedure resulted in less biased estimates in the face of both guessing and “careless” responses.

Barton and Lord (1981) were also concerned that the 3PM may excessively punish errors by high-ability students. They explored whether changing the upper asymptote improved scoring on standardized tests. They added a fourth parameter, d , to drop the upper asymptote below 1:

$$P_j(\theta) = c_j + (d - c_j) \frac{e^{1.7a_j(\theta - b_j)}}{1 + e^{1.7a_j(\theta - b_j)}} \quad (6)$$

Barton and Lord then reestimated test scores for thousands of students who had taken the Scholastic Aptitude Test (SAT), Graduate Record Examination (GRE), and Advanced Placement

Figure 2
 Posterior Distributions for Three-Item Test With Easiest
 Item Answered Incorrectly



Note. For the solid line, question difficulties are given by $b_j = [-1, 0, 1]$ (same as improper likelihood in Figure 1). For the dashed line, question difficulties are given by $b_j = [-1, 2.5, 2.5]$. The posterior mean is *lower* and the posterior variance is *smaller* when these two harder items are answered correctly.

(AP) exams to determine the effect of fixing d at 0.99 and 0.98. They concluded that changes in ability estimates were too small to be of practical significance, especially given the difficulty (at that time) of implementing the new model.

Both these examples assume that final estimates are derived after the student has completed a static test in which all students receive predetermined items from throughout the entire ability range. In CATs, however, estimation is dynamic and items are selected based on accumulating information regarding student performance. Thus, unlike traditional test scoring, early aberrant responses cannot be discounted with a retrospective evaluation of the entire response vector, as the early answers provide the only information with which to continue the CAT and select future items. Although updating $\hat{\theta}$ after each response makes CATs very efficient, it may be problematic for high-ability students who miss initial questions. In such cases, the (almost-all) correct responses to easy items that are far from the respondent's true θ contribute little (or perhaps negative) information, resulting in a very slow climb. Modifying the 3PM may facilitate faster recovery of the algorithm if the student makes early mistakes.

The authors argue that it is worth reconsidering Barton and Lord's (1981) 4PM for use in CATs. The upper asymptote <1 allows a small probability of error even by very high-ability students, reducing the asymmetry of the 3PM. This might have a more obvious impact on test scoring in the early stages of a CAT, when relatively few items have been answered.

Reducing the Impact of Early Mistakes in CATs

The authors revisit the three-item example and consider model adjustments that might reduce the impact of early mistakes by high-ability students. In Figure 3a, the posterior distribution is shown for the response pattern $x = (0, 1, 1)$, where $b_2 = b_3 = 2.5$. The left panel shows the 3PM5. The left panel shows the 3PM and the right panel shows Barton and Lord's (1981) 4PM with $d = 0.98$. The posterior distribution for the 4PM is similar to that of the 3PM, but the density above $y = 1$ is slightly greater, and the posterior mean is higher, $\hat{\theta} = -0.73$

(compared with $\hat{\theta} = -0.91$). When two more items with $b_4 = b_5 = 2.5$ are added, the 3PM barely moves (Figure 3b, left), but a second mode is evident in the 4PM (right). The posterior distribution for the 4PM now allows the possibility that the first response was aberrant, and that the four correct responses to difficult items reflect the true θ . The two modes in the posterior distributions represent opposing hypotheses: Either the student is truly of low ability and has just been lucky on the difficult items, or the student is of high ability and was unlucky on the easy item. Even after adding two more difficult items, $b_6 = b_7 = 2.5$, the 3PM posterior distribution still barely acknowledges the second hypothesis, but it becomes the dominant mode for the 4PM posterior distribution. Clearly, the 4PM seems better able to accommodate the possibility that a high-ability student carelessly missed an easy item.

Another modification that might make the model more flexible to aberrant responses is to impose a less informative prior distribution. Ordinarily, the prior is set to $p(\theta) \sim N(0,1)$ because θ is assumed to follow the standard normal distribution in the population (Bock & Mislevy, 1982; Mislevy, 1984; Owen, 1975; Wang & Vispoel, 1998). Nevertheless, in a CAT where estimation is continuous and begins after the first item, it is possible that using a less informative prior, such as $p(\theta) \sim N(0,2)$, would allow $\hat{\theta}$ to ascend more quickly and further improve estimation.

When the example in Figure 3 was reconsidered with a 4PM and $p(\theta) \sim N(0,2)$, a similar contrast between the 3PM and 4PM was found. With the less informative prior, however, the 4PM adapted more quickly than it did with the standard normal prior. The second mode appeared after only two items with $b = 2.5$ and became the dominant mode after only four items with $b = 2.5$. In the context of early aberrant responses, the less informative prior allows the high-ability student to fall more quickly after early misses, but it also allows $\hat{\theta}$ to rise faster.

Present Study

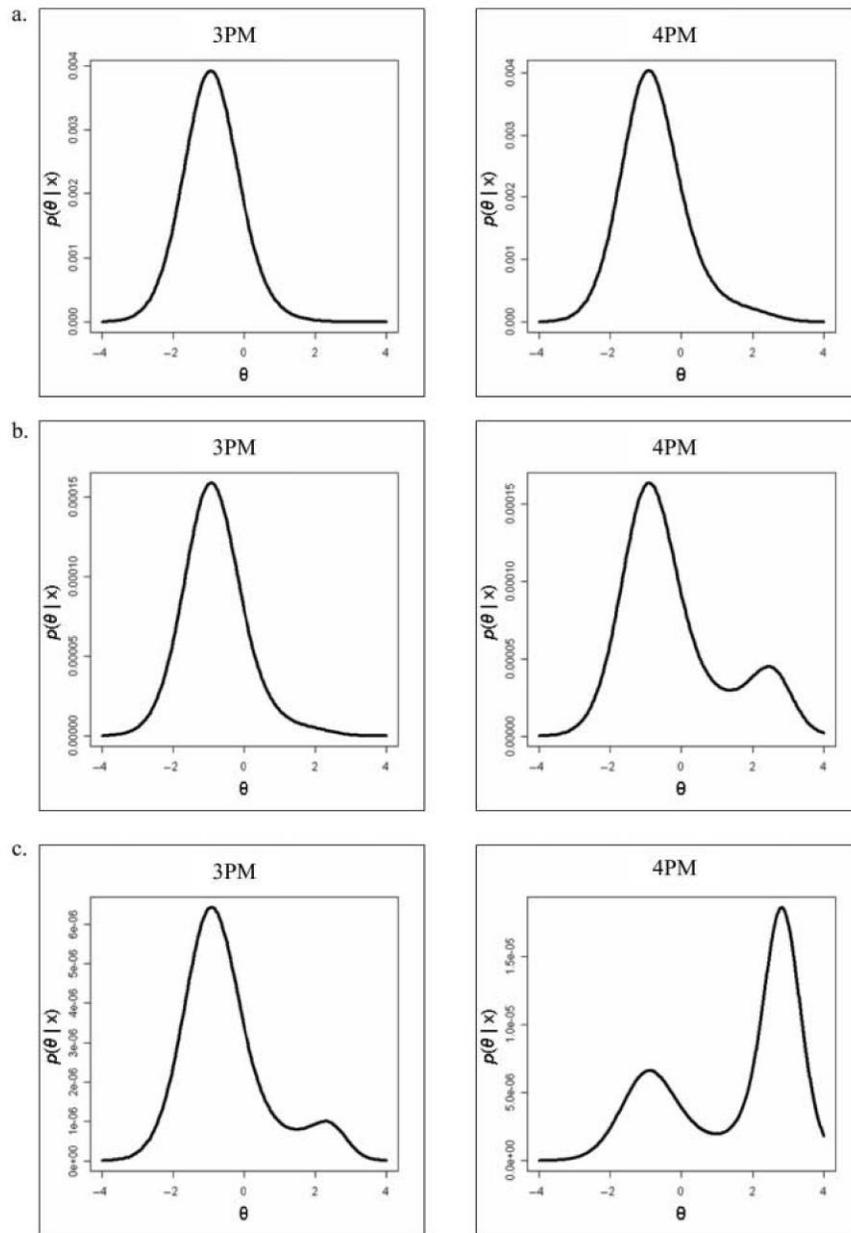
The following analyses explore the extent to which early aberrant answers by high-ability students can lead to underestimation in CAT, and whether the impact of early mistakes can be reduced. In this study it is hypothesized that (a) a typical CAT 3PM algorithm will ascend slowly after starting with incorrect early answers, regardless of estimation method; (b) the bias will be less serious when a low-ability student gets lucky and answers the first two items correctly, but finishes the test by answering the remaining items according to true θ ; (c) the 4PM proposed by Barton and Lord (1981) will have utility in CAT by greatly reducing the risk for biased estimates; (d) the effectiveness of the 4PM will be further strengthened by using a less informative prior than the standard normal; and (e) the 4PM will outperform an alternative approach in which less discriminating items are selected early in a CAT to reduce initial step size.

CAT Simulation I

Data and Method

Simulation. One hundred samples of 5,000 students were simulated, with $\theta \sim N(1,0)$. Within each sample, students were rank ordered, and the top and bottom 10% ($N=1,000$ total) were

Figure 3
The Posterior Distribution for the Three-Item Test in Figure 2
Under 3PM and 4PM After Additional Items Are Answered Correctly



Note. (a) Left: The posterior distribution for the three-item test in Figure 2 (dashed curve; $b_j = [-1, 2.5, 2.5]$) under the three-parameter model (3PM). Right: The posterior distribution for the same response pattern under the four-parameter model (4PM). Panels (b) and (c) represent the posterior distribution after two and four additional items with $b_j = 2.5$ are answered correctly. The final panel represents $\mathbf{x} = [0, 1, 1, 1, 1, 1, 1]$ for items $b_j = [-1, 2.5, 2.5, 2.5, 2.5, 2.5, 2.5]$.

selected. An idealized item pool of 1,000 items was created, with $a_i \sim N(1.1, 0.1)$, $b_i \sim U(-3, 3)$, and $c_i \sim N(0.2, 0.04)$. Ignoring security issues or content balancing, this item pool should represent a highly efficient testing environment.

In the simulated CAT, the initial ability estimate, $\hat{\theta}^0$, was set to 0 for each student. Items were selected using a simple maximum-information criterion. Specifically, an information grid was constructed with items rank ordered for their Fisher's information value at discrete increments of 0.05 between -4 and 4. Fisher's information for the 3PM is given by

$$I_j(\theta) = \frac{(1.7a_j)^2(1 - c_j)}{(c_j + e^{1.7a_j(\theta - b_j)})(1 + e^{-1.7a_j(\theta - b_j)})^2}. \quad (7)$$

and Fisher's information function for the 4PM is given by

$$I_j(\theta) = \frac{(1.7a_j)^2(d - c_j)^2}{(c_j + de^{1.7a_j(\theta - b_j)})(1 - c_j + (1 - d)e^{1.7a_j(\theta - b_j)})(1 + e^{-1.7a_j(\theta - b_j)})^2}. \quad (8)$$

The item with the most information at $\hat{\theta}^{k-1}$ that had not yet been answered was selected and administered, where k was the current item number. Under standard conditions, the simulated response x_k was correct with $P_j(\theta)$ determined by the item's parameters.

Tests with 15, 30, and 45 items were simulated under three conditions. In the first θ ("standard"), the full sample of students answered according to their true y for all items. In the second condition, students in the top and bottom 10% of the distribution answered the first two items correctly, regardless of their true y , and then answered the remaining items according to their true ability. In the third condition, students in the top and bottom 10% of the distribution answered the first two items incorrectly, and then answered the remaining items according to their true y : Student ability, $\hat{\theta}^k$, was estimated in one of three ways: (a) as the \sim , (b) as the EAP estimate, with $p(\theta) \sim N(0,1)$, and (c) as the EAP estimate, with $p(\theta) \sim N(0,2)$. Results were obtained first for the 4PM and then repeated for the \sim , in which d was fixed at 0.98.

Evaluation. Within each replication, average bias and root mean square error (RMSE) were computed separately for the top and bottom 10% samples:

$$\text{Bias}(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^N (\hat{\theta}_i - \theta_i). \quad (9)$$

$$\text{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (\hat{\theta}_i - \theta_i)^2}. \quad (10)$$

Bias and RMSE were then averaged across the 100 replications. In addition, coverage was computed as the percentage of cases in which the true y fell within the 95% confidence intervals, and these percentages were averaged across the 500 replications. The \sim intervals were generated by calculating the total test information at the conclusion of the test and constructing an interval \pm two standard errors from the final $\hat{\theta}$. This method assumes a quadratic approximation at the posterior mean, which would not be appropriate early in the test but should be less problematic for the final estimate after several items.

Table 1
Bias, Root Mean Square Error (RMSE), and Percentage Coverage at the
Completion of a Computerized Adaptive Test: Three-Parameter Model (3PM)

	Bias			RMSE			Percentage Coverage of 95% Confidence Intervals		
	ML	$N(0,1)$	$N(0,2)$	ML	$N(0,1)$	$N(0,2)$	ML	$N(0,1)$	$N(0,2)$
	Standard performance								
Full sample									
15 items	0.05	0.00	0.00	0.42	0.32	0.34	0.92	0.95	0.95
30 items	0.01	0.00	0.00	0.25	0.22	0.23	0.94	0.95	0.95
45 items	0.006	0.00	0.00	0.19	0.18	0.19	0.94	0.95	0.95
Miss first two items									
Top 10%									
15 items	-2.17	-1.29	-1.83	2.21	1.35	1.88	0.00	0.02	0.00
30 items	-0.71	-0.56	-0.67	0.80	0.64	0.75	0.47	0.51	0.50
45 items	-0.28	-0.31	-0.32	0.38	0.39	0.40	0.77	0.70	0.75
Bottom 10%									
15 items	-0.19	0.13	-0.08	0.37	0.32	0.30	0.91	0.94	0.96
30 items	-0.08	0.07	-0.04	0.24	0.23	0.22	0.93	0.94	0.96
45 items	-0.06	0.05	-0.03	0.20	0.19	0.18	0.94	0.94	0.95
Get first two items correct									
Top 10%									
15 items	0.08	-0.17	-0.02	0.40	0.35	0.32	0.91	0.93	0.96
30 items	0.03	-0.08	-0.01	0.25	0.24	0.22	0.93	0.94	0.96
45 items	0.02	-0.06	-0.008	0.20	0.19	0.19	0.94	0.95	0.95
Bottom 10%									
15 items	0.53	0.49	0.24	1.30	0.72	0.67	0.70	0.76	0.89
30 items	0.08	0.17	0.06	0.43	0.33	0.30	0.90	0.89	0.94
45 items	0.03	0.10	0.04	0.24	0.24	0.22	0.93	0.92	0.94

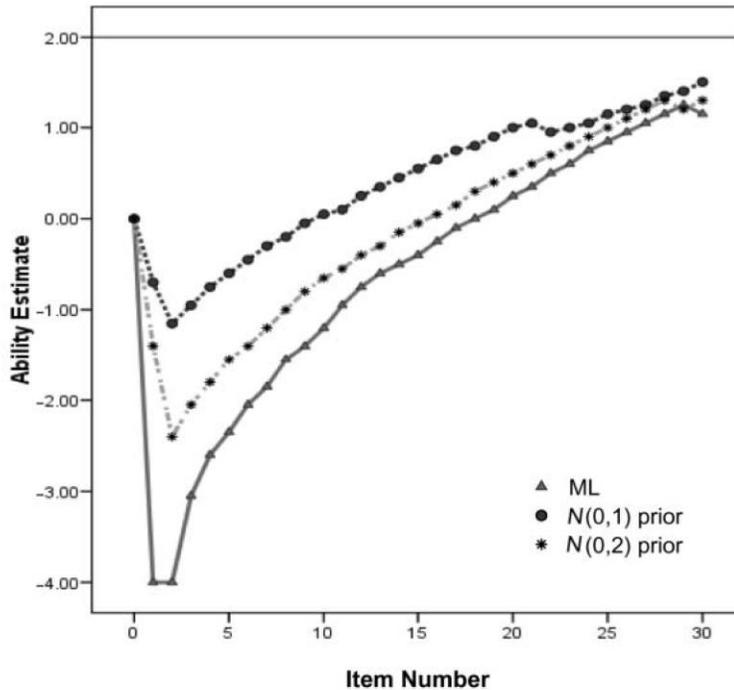
Note. ML = maximum likelihood.

Results

Three-parameter model (3PM). Under standard performance, y in the full sample was well-estimated with our simulated CAT. When students performed according to true y throughout the test, there was minimal bias ($M = 0.00 - 0.05$) under MLE and no bias under Bayesian estimation (Table 1). Coverage of the confidence intervals was at or just under 95% with all three estimation procedures. RMSE ranged from .32 for the 15-item test to .18 for the 45-item test when $p(\theta) \sim N(0,1)$, and was slightly higher (0.19-0.42) under MLE.

When students missed the first two items and then performed according to their true y for the remainder of the test (second condition), estimates for the top 10% of students were strongly negatively biased. On the 30-item test, mean (M) bias ranged from 0.71 standard deviation (SD) below the true θ under MLE to 0.56 SD below the true y when $p(\theta) \sim N(0,1)$. Even after 45 items, $\hat{\theta}$ was still about 1/3 SD too low across all three estimation methods. Not only was $\hat{\theta}$ biased, but interval coverage was poor, with almost no coverage on the 15-item test ($M < 0.02$), and only 70% to 77% coverage for the 45-item test. Predictably, for the bottom 10% of students there was considerably less bias and RMSE was smaller, especially on longer tests. On the 30-item test, for example, there

Figure 4
 Ability Estimates Updated After Each Item in a
 Three-Parameter Model Computerized Adaptive Test



Note. The lines represent the estimation pattern after two incorrect answers followed by true performance. After 30 items, $\hat{\theta}$ does not reach the true θ of 2.0.

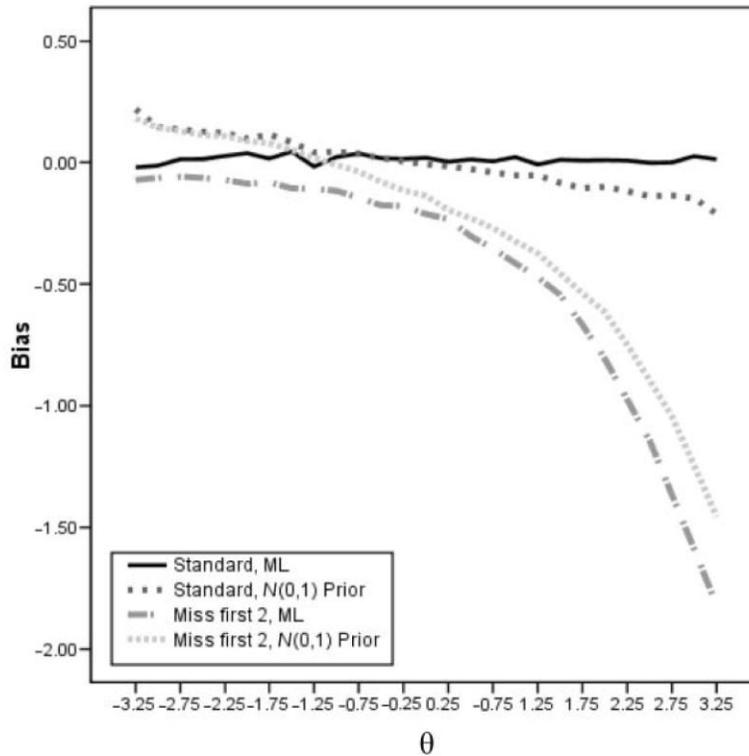
was a small negative bias for the MLE ($M = -0.08$) and for the EAP with the less informative prior ($M = -0.04$). There was a small positive bias ($M = 0.07$) under the standard normal prior because the prior tends to pull extreme values of $\hat{\theta}$ toward the mean. Coverage was over 90% regardless of test length or estimation method.

In the third condition, when students started the test with two correct answers, bias was much less pronounced. For high-ability students, there was a small positive bias for the MLE and a small negative bias for EAP, but RMSE and coverage were similar to the standard condition, especially on longer tests. Low-ability students, who would be the ones to benefit from a good start to the test, had positive bias on shorter tests ($M = 0.53$ for the MLE to $M = 0.24$ for EAP on the 15-item test), but on longer tests, they were estimated almost as accurately as in the standard condition ($M = 0.06 - 0.17$ on the 30-item test and $M = 0.03 - 0.10$ on the 45-item test).

To illustrate the negative bias when a student misses the first two items, the trajectory of estimates for a single high-ability student (true $\theta = 2$) across a 30-item test is plotted in Figure 4. Missing the first two items resulted in a considerable initial drop that was largest under MLE ($\hat{\theta} = -4.0$) and smallest under the standard normal prior ($\hat{\theta} = -1.15$). The initial drop was followed by a very slow ascent in $\hat{\theta}$, such that by item 30, the true θ was still not reached by any of the estimation procedures, even though 27 of the first 30 items were answered correctly in each case.

To examine the degree of risk across ability levels, 500 students were simulated for y between -3.25 and 3.25 , in increments of 0.25 , using MLE and the standard normal prior. (When $p(\theta) \sim N(0,2)$,

Figure 5
 Bias at the End of a 30-Item Computerized Adaptive Test
 as a Function of True Ability (Three-Parameter Model)



Note. Mean bias in the final $\hat{\theta}$ after 30 items under two conditions: student always answers according to true θ (standard performance) and student misses the first two items, regardless of true θ .

, results were nearly identical to those using MLE.) The mean bias was plotted at the end of a 30-item test for these individuals in the standard condition and after missing the first two items on the test (Figure 5). When the first two items were missed, there was underestimation greater than 0.20 SD for students with $\theta = 0$ in both estimation procedures. The bias obviously increased for students with $\theta > 0$.

In sum, under the 3PM, early mistakes by high-ability students lead to negatively biased final estimates, particularly on shorter tests. The opposite response pattern, in which low-ability students essentially start with lucky guesses, resulted in much less bias. As expected, the CAT algorithm can recover downward more quickly than it can recover upward. The authors had argued earlier that this was most likely due to the assumption built into the 3PM that the upper asymptote is 1.

Four-parameter model (4PM). Whether the bias that occurred for high-ability students who missed early items could be reduced under the 4PM, with $d = 0.98$, was next evaluated. Table 2 shows that under standard performance conditions, there was a small positive bias in the full sample ($M = 0.03$ for both priors; $M = 0.04 - 0.07$ for MLE) but RMSE and interval coverage were essentially unchanged from the standard 3PM. Under the condition of early aberrant responses (i.e., missing the first two items), estimates were greatly improved for high-ability examinees. On

Table 2
Bias, Root Mean Square Error (RMSE), and Percentage Coverage at the Completion of a Computerized Adaptive Test: Four-Parameter Model (4PM)

	Bias			RMSE			Percentage Coverage of 95% Confidence Intervals		
	ML	$N(0,1)$	$N(0,2)$	ML	$N(0,1)$	$N(0,2)$	ML	$N(0,1)$	$N(0,2)$
	Standard performance								
Full sample									
15 items	0.07	0.03	0.03	0.42	0.32	0.34	0.93	0.96	0.96
30 items	0.04	0.03	0.03	0.25	0.23	0.23	0.94	0.95	0.95
45 items	0.04	0.03	0.03	0.20	0.19	0.19	0.95	0.95	0.95
Miss first two items									
Top 10%									
15 items	-1.26	-0.92	-0.48	1.33	1.12	0.77	0.47	0.60	0.88
30 items	-0.03	-0.22	-0.08	0.38	0.37	0.33	0.92	0.90	0.95
45 items	-0.002	-0.10	-0.03	0.25	0.24	0.23	0.94	0.93	0.95
Bottom 10%									
15 items	-0.11	0.15	-0.04	0.37	0.33	0.30	0.93	0.94	0.97
30 items	-0.03	0.10	-0.006	0.24	0.24	0.22	0.95	0.94	0.96
45 items	-0.01	0.07	0.004	0.19	0.20	0.18	0.95	0.94	0.96

Note. ML = maximum likelihood.

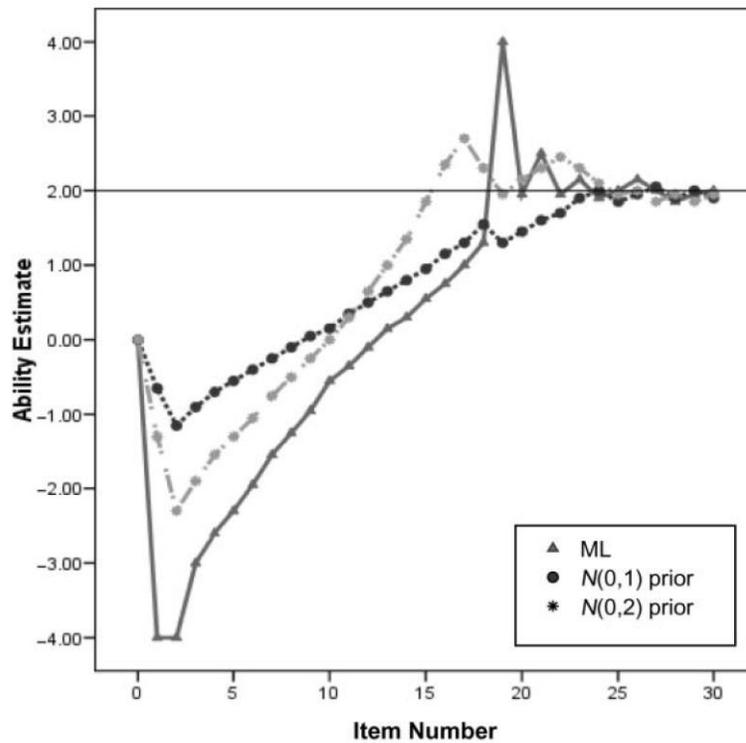
the 30-item test, for example, the bias for high-ability students was -0.22 for the standard normal prior (compared with -0.56 for the 3PM), and 0.56 for the 3PM, and -0.03 for the MLE (compared with -0.71 for the 3PM). RMSE was at most 0.38 (compared with 0.80 for the 3PM) and the coverage of the confidence intervals was 90% or more under all three estimation procedures. After 45 items, mean bias had been reduced to -0.10 when $p(\theta) \sim N(0,1)$ and to -0.002 for ML estimation and coverage had improved to 93%-95%. For students in the bottom 10% of the distribution, bias on the 45-item test with the standard normal prior increased from 0.05 to 0.07 , but RMSE and coverage were essentially the same as for the 3PM.

In Figure 6, the trace plot for a high-ability student who missed the first two items was revisited. The errors caused an initial drop almost identical to the drop in the 3PM, but $\hat{\theta}$ ascended faster because the upper asymptote of 0.98 discounted the early mistakes. When $p(\theta) \sim N(0,2)$, $\hat{\theta}$ reaches 2.0 (the true θ) by item 16. By item 24, $\hat{\theta}$ has reached 2.0 under all three estimation procedures. Then, as expected, $\hat{\theta}$ continued to oscillate a little above and below 2.0 until the end of the test, because the student was encountering items that closely matched his or her true y .

Next, 500 students were again simulated at each y between -3.25 and 3.25 in increments of 0.25 . Mean bias at the end of a 30-item test is plotted in Figure 7. In the standard condition, bias 0.25 . Mean bias at the end of a 30-item test is plotted in Figure 7. In the standard condition, bias using the 4PM was slightly higher across the entire y range than it was for the 3PM. However, under conditions where the first two items are missed, estimation had improved considerably. Bias never exceeds $|0.10|$ under MLE and remains below $|0.40|$ under the standard normal (part of this bias reflects the pull toward the mean by the prior).

In sum, the 3PM can produce negatively biased estimates for high-ability students who make mistakes early in a CAT. The bias is not symmetric, as the algorithm descends after a lucky start

Figure 6
 Ability Estimates Updated After Each Item
 in a Four-Parameter Model Computerized Adaptive Test



Note. By item 24, $\hat{\theta}$ reaches the true θ of 2.0 for all three estimation procedures.

much faster than it ascends from an unlucky start. As predicted, the 4PM reduced the negative bias much faster than it ascends from an unlucky start. As predicted, the 4PM reduced the negative bias considerably, and bias was further reduced when a less informative prior was used.

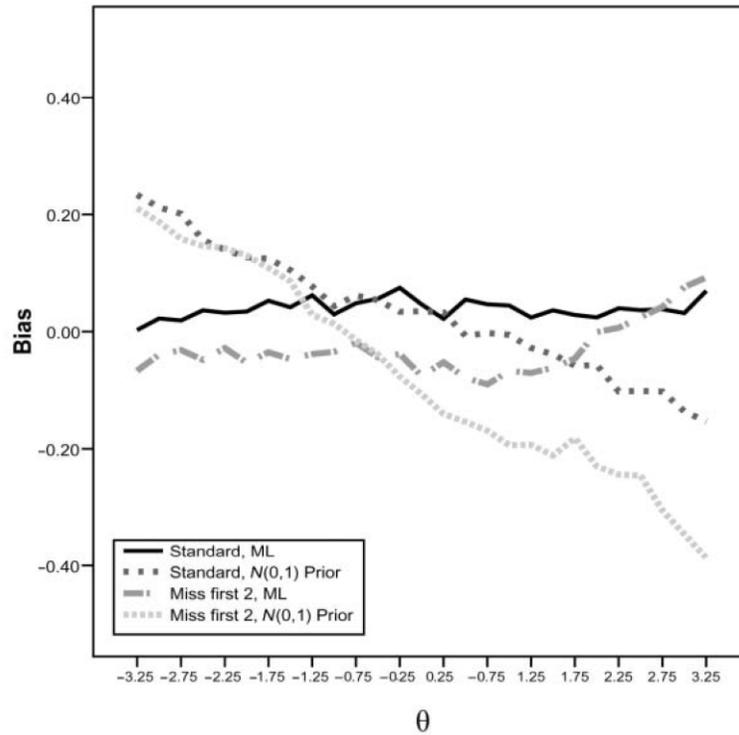
As discussed in the introduction, recent CAT research has explored the strategy of stratifying the item pool so that less discriminating items are used first. The main justification is to retain items with higher a values for use later in the test. However, it may also be true that starting the test with less discriminating items could reduce the bias due to early errors by high-ability students; the lower a ensures a wider ability range across which there is at least a modest chance of getting the item wrong. Therefore a second simulation was conducted to evaluate the impact of using items with lower a parameters early in the test and to explore how this approach compared with estimation under the 4PM.

CAT Simulation II

Data and Method

Simulation. The second simulation followed the same procedure described earlier, using the same 100 samples of students generated in Simulation I. Performances on 15-, 30-, and 45-item CATs using three estimation methods were assessed first for the 3PM and then for the 4PM. Across

Figure 7
 Bias at the End of a 30-Item Computerized Adaptive Test
 as a Function of True Ability (Four-Parameter Model)



Note. Mean bias in the final $\hat{\theta}$ after 30 items, under two conditions: student always answers according to true θ (standard performance) and student misses the first two items, regardless of true θ .

all conditions, the item-selection algorithm chose the item with the largest expected information at θ^{k-1} . The only modification was that the discrimination parameter for the first two items was fixed to $a = 0.9$, such that the discrimination of these two items was set considerably lower than for the rest of the items in the test. This condition is referred to as the fixed-a condition, and the Simulation I results are referred to as the original 3PM and the original 4PM results.

Results

Three-parameter model (3PM). Under standard performance, results for the fixed-a condition were identical to the original 3PM for the full sample (Table 3). When students in the top 10% of were identical to the original 3PM for the full sample (Table 3). When students in the top 10% of the distribution missed the first two items, however, performance was better in the fixed-a condition than the original 3PM but worse than the original 4PM. For example, on a 30-item test with MLE, bias was -0.39 for the fixed-a condition, compared with -0.71 for the original 3PM and -0.03 for the original 4PM. Similarly, RMSE was 0.52 for the fixed-a condition, compared with 0.8 for the original 3PM and 0.38 for the original 4PM. Although coverage exceeded 90% for the original 4PM with tests of 30 items or more, coverage was still only 84% to 89% after 45 items in the fixed-a condition. The coverage, however, was better than for the original 3PM (70% – 77%).

Table 3
 Bias, Root Mean Square Error (RMSE), and Percentage Coverage at the Completion of a Computerized Adaptive Test: Lower α -Parameter on First Two Items

	Bias			RMSE			Percentage Coverage of 95% Confidence Intervals		
	ML	$N(0,1)$	$N(0,2)$	ML	$N(0,1)$	$N(0,2)$	ML	$N(0,1)$	$N(0,2)$
	Standard performance: 3PM								
Full sample									
15 items	0.05	0.00	0.00	0.41	0.32	0.34	0.92	0.95	0.96
30 items	0.01	0.00	0.00	0.25	0.22	0.23	0.94	0.95	0.95
45 items	0.008	0.00	0.00	0.19	0.18	0.19	0.94	0.95	0.95
Miss first two items: 3PM									
Top 10%									
15 items	-1.78	-0.89	-1.04	1.83	0.98	1.12	0.00	0.41	0.47
30 items	-0.39	-0.34	-0.30	0.52	0.44	0.41	0.76	0.75	0.83
45 items	-0.17	-0.20	-0.16	0.29	0.29	0.27	0.88	0.84	0.89
Bottom 10%									
15 items	-0.14	0.12	-0.04	0.36	0.33	0.31	0.93	0.94	0.96
30 items	-0.07	0.07	-0.02	0.24	0.23	0.23	0.94	0.94	0.95
45 items	-0.04	0.05	-0.01	0.19	0.19	0.19	0.94	0.95	0.95
Miss first two items: 4PM									
Top 10%									
15 items	-1.01	-0.71	-0.46	1.10	0.81	0.69	0.72	0.66	0.88
30 items	-0.05	-0.21	-0.10	0.34	0.34	0.31	0.94	0.89	0.94
45 items	-0.01	-0.11	-0.04	0.23	0.23	0.22	0.95	0.93	0.95
Bottom 10%									
15 items	-0.09	0.16	0.01	0.35	0.35	0.31	0.94	0.93	0.97
30 items	-0.02	0.10	0.02	0.24	0.24	0.23	0.95	0.94	0.96
45 items	-0.01	0.08	0.02	0.19	0.20	0.19	0.95	0.94	0.96

Note. ML = maximum likelihood; 3PM = three-parameter model; 4PM = four-parameter model. α -parameter fixed at 0.9 for first two items.

For students in the bottom 10% of the distribution, bias was generally similar or slightly smaller for the fixed- α condition than for the original 3PM and 4PM with the standard normal prior, but for MLE and the less informative prior, it was between the original 3PM and 4PM. RMSE and interval coverage were nearly identical across all three models and all three estimation procedures.

Four-parameter model (4PM). Finally, when the fixed- α condition was used with the 4PM, bias and coverage for students in the top 10% of the distribution improved for the 15-item test but were essentially identical to the original 4PM on the longer tests. RMSE was smallest with the combined fixed- α and 4PM approach under all test lengths, particularly on the 15-item test.

Discussion

There is a popular perception that the initial items on a CAT are especially influential in determining a student's final score. Some commercial test preparation services even recommend spending extra time on the first few items to ensure the best possible score (Kaplan, 2004; Lurie, Pecsénye, Robinson, & Ragsdale, 2005). Although this advice reflects a general misconception about how a CAT functions, the results of this study show that the advice does contain a grain of truth. Under the 3PM, the final $\hat{\theta}$ was strongly biased for high-ability students who underperformed early in the CAT. After missing the first two items, high-ability students were unable to ascend to their true θ , even after 45 items. However, spending extra time on the initial items is unlikely to help average or low-ability students obtain higher scores, especially when trade-offs regarding allocation of time are taken into account. This study showed that an unexpectedly good start was mostly erased as the CAT algorithm descended more quickly to a final $\hat{\theta}$ more reflective of true ability.

The problem that $\hat{\theta}$ sometimes has a large initial drop and then ascends continuously until the end of a CAT is recognized by some testing professionals, but has received very little attention in the literature (Chang, 2004). Regarding biased estimates more generally, Chang and Ying (2002) and Chang (2004), argued that smaller step sizes later in a CAT (after the best items have been used) prevent students who over- or underperform on the initial items from reaching their true θ . Altering an item-selection algorithm so that it does not use items with the highest a values first can limit overexposure and improve θ estimation (Chang & Ying, 1999).

The results of this study, however, showed that underestimation bias was not only a consequence of shrinking step sizes. Although fixing the value of a to 0.9 for the first two items reduced the impact of early aberrant responses, the authors showed that because of the 3PM's assumptions, the impact of early aberrant responses, the authors showed that because of the 3PM's assumptions, there was a lingering effect of early aberrant responses not corrected by increasing step size alone. In addition, the three-item example of this study showed that administering more difficult items to a student who misses an easy item can actually lead to a lower a student who misses an easy item can actually lead to a lower $\hat{\theta}$. Chang (2004) further speculates that the step-size argument also implies the potential for overestimation after a lucky start. However, the authors have shown that $\hat{\theta}$ falls much faster than it rises and argued that the asymmetry occurred because the lower asymptote of c_j can accommodate lucky guesses by low-ability students, whereas the upper asymptote of 1 cannot accommodate unlucky mistakes by high-ability students. The problem of bias because of aberrant early responses is much more serious when high-ability students do poorly early in the test.

Two model adjustments that can allow the algorithm to ascend more quickly after early mistakes by high-ability students were suggested. Barton and Lord's (1981) 4PM was revisited, arguing that in the context of dynamic θ estimation, the 4PM may be better than the 3PM. The authors showed that setting the upper asymptote slightly below 1 and using a less informative prior considerably reduced bias for high-ability students who missed the first two items. The model adjustments did not compromise estimation quality under standard performance conditions.

Practical Implications

So far, a statistical solution to a potential estimation problem has only been considered. But what is the "real world" significance? How often do high-ability students miss early items? Is it appropriate to adjust the model to accommodate such mistakes? Although the model predicts that a high-ability student should very rarely miss the first two items on a test that starts with average items, there are plausible reasons for early poor performance. Nervousness, unfamiliarity with the testing situation, distractions, unexpected content, and carelessness are all possible reasons for early unexpected errors. The prevalence of such behaviors is not known, but in 2000, ETS allowed 0.5% of examinees to retest in the context of estimation problems in the GRE CAT (Carlson, 2000). Although the reason for allowing 1 in 200 people to retake the test is unknown, Chang (2004) argued that many of these students likely had problematic response patterns that suggested their final scores were underestimated.

If high-ability students underperform on the first few items of a CAT, is it appropriate to adjust the model to anticipate such mistakes, or should their responses be "punished" with a lower score? CATs are supposed to be flexible and dynamic so as to quickly arrive at accurate estimates. If high-ability students return to their true level of performance for the rest of the test, there should not be a built-in model feature that prevents full recovery. If it is possible to accommodate early aberrant responses without substantially altering the measurement properties of the test under normal performance, then such model adjustments deserve consideration.

Exactly how often high-ability students begin a test with unexpectedly poor performance is an empirical question. In a high-stakes setting, students are motivated to maximize their performance, so the actual frequency

may below. However, this study showed that a large percentage of test takers are vulnerable to underestimation, should they get off to an uncharacteristically bad start. In fact, the analyses of this study indicated that most students above $\theta = 0$ are at risk. Model adjustments could therefore act as an insurance, protecting the interests of test takers and test administrators.

The findings of this study are also relevant to uses of IRT in domains other than academic testing. For example, in fields such as psychopathology and personality assessment, IRT is becoming increasingly popular. In such applications, respondents at the low and high end of the trait distribution may provide aberrant responses for reasons, such as social desirability, lying, multidimensionality, or simply because not all symptoms are universally present (or universally absent) at the poles of the distribution (Reise & Waller, 2003; Waller & Reise, in press). Models must accommodate the true response pattern in the tails of the distribution or risk providing biased estimates of θ . In addition, item calibration is complicated when questionable assumptions are made in the tails of the distribution (see Rouse, Finger, & Butcher, 1999, for an example with a psychoticism scale). More generally, the measurement model must be aligned with actual response patterns, whether the model is used for dynamic assessment in CATs, or in the fixed-length instruments widely employed in psychological research and practice.

Limitations

Although the authors identified several potential model adjustments that can improve a known problem with θ estimation, they noted some limitations to their work. First, the simulated CAT is not an exact replica of any operational CAT. Actual testing algorithms may incorporate ad hoc procedures early in the test to mitigate the effects described here. Second, the simulations in this study were done under ideal conditions with a rich database of questions and no constraints about test security, item exposure, or content balancing. The measurement properties of the CAT could be different in the presence of these real-world constraints (Chen, Ankenmann, & Chang, 2000; Chang & Ying, 1999). Alternative approaches to Fisher's information for item selection also were not considered. It could be that the early aberrant performance is less influential under an item selection algorithm that considers expected information across the likelihood (Chang & Ying, 1996; Cheng & Liou, 2000; van der Linden, 1998).

Finally, the authors acknowledge that much validation research would need to be carried out before widely implementing the 4PM for CAT. The purpose of this article was to document an estimation problem in CATs and give a promising solution. Many theoretical and empirical issues must be investigated before making wholesale changes to standard CAT procedures.

By a similar token, the fixed- a simulation of this study was not a full treatment of the stratified- a approach. Typically, the stratification is done across the test, and not just for the first two items.

The authors believe that here, too, there is much research to be done in terms of evaluating overall test properties and interactions with other aberrant response patterns. For example, would the detrimental effects of guessing at the end of the test be exacerbated when the stratification has saved some of the most discriminating items until the end?

Conclusion

This study has shown that underestimation can occur in a CAT because of early underperformance by otherwise high-ability students and shows why the 3PM is quicker to descend than it is to ascend. This study also shows that using Barton and Lord's (1981) 4PM and using a less informative prior both reduce the bias at the end of a CAT after initial mistakes. Further research should be undertaken to investigate the consequences of implementing model adjustments in real testing situations to avoid underestimation.

References

Baker, F. B., & Kim, S. (2004). *Item Response Theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

- Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model (Tech. Rep. RR-81-20). Princeton, NJ: Educational Testing Service.
- Bock, R., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment, *Applied Psychological Measurement*, 6, 431-444.
- Bradlow, E. T. (1996). Negative information and the Three-Parameter Logistic Model. *Journal of Educational and Behavioral Statistics*, 21, 179-185.
- Carlson, S. (2000). ETS finds flaws in the way online GRE rates some students. *Chronicle of Higher Education*, 47, a47.
- Chang, H. H. (2004). Understanding computerized adaptive testing: From Robbins-Munro to Lord and beyond. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 117-133). New York: Sage.
- 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. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222.
- Chang, H. H., & Ying, Z. (2002). To weight or not to weight—Balancing influence of initial and later items in adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Chen, S. Y., Ankenmann, R. D., & Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, 24, 241-255.
- Cheng, P. E., & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement*, 24, 257-265.
- Kaplan Publishing. (2004). GRE exam, 2005 edition. New York: Simon & Schuster.
- Lurie, K., Pecsénye, M., Robinson, A., & Ragsdale, D. (2005). *Cracking the GRE*. New York: Princeton Review.
- Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, 23, 187-194.
- Mislevy, R. J. (1984). Estimating latent distributions, *Psychometrika*, 49, 359-381.
- Mislevy, R. J., & Bock, R. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement*, 42, 725-737.
- Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of American Statistical Association*, 70, 351-356.
- Passos, V. L., Berger, M. P. F., & Tan, F. E. (2007). Test design optimization in CAT early stage with the nominal response model. *Applied Psychological Measurement*, 31, 213-232.
- Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, 8, 164-184.
- Rouse, S. V., Finger, M. S., & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2
Downloaded from
PSY-5 scales. *Journal of Personality Assessment*, 72, 282-307.
- van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63, 201-216.
- van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detecting person-misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201-219). Boston: Kluwer Academic.
- Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, 35, 109-135.
- Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., et al. (2000). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.
- Waller, N. G., & Reise, S. P. (in press). Measuring psychopathology with nonstandard IRT models: Fitting the four parameter model to the MMPI. In S. Embretson & J. S. Roberts (Eds.), *New directions in psychological measurement with model-based approaches*.

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

Acknowledgments

Support for this research was provided by NSF award SES-0352191 (PI Loken) and the National Institute on Drug Abuse (DA 017629; DA 024497-01).