Modeling Student Test-Taking Motivation in the Context of an Adaptive Achievement Test

Steven L. Wise

Northwest Evaluation Association

G. Gage Kingsbury

Psychometric Consultant

Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April, 2015. The views expressed in this paper are solely those of the authors and they do not necessarily reflect the positions of NWEA. Correspondence concerning this paper should be addressed to Steven L. Wise, Northwest Evaluation Association, 121 NW Everett St, Portland, Oregon 97209. Email: steve.wise@nwea.org.

Abstract

This study examined the utility of response time-based analyses in understanding the behavior of unmotivated test takers. For an adaptive achievement test, patterns of observed rapid-guessing behavior and item response accuracy were compared to the behavior expected under several types of models that have been proposed to represent unmotivated test taking behavior. Test taker behavior was found to be inconsistent with these models, with the exception of the effort-moderated model (S.L. Wise & DeMars, 2006). Effort-moderated scoring was found to both yield scores that were more accurate than those found under traditional scoring, and exhibit improved person fit statistics. In addition, an effort-guided adaptive test was proposed and shown to alleviate item difficulty mis-targeting caused by unmotivated test taking.

Modeling Student Test-Taking Motivation in the Context of an Adaptive Achievement Test

Score validity is diminished by unmotivated test taking. This simple statement has strong implications for achievement testing, because a universal assumption underlying currently-used psychometric models is that whenever we administer a test item, a test taker will be motivated to put forth the effort needed to select or produce the correct answer. The assumption, however, is frequently violated—particularly with low-stakes tests, for which test takers often perceive that there are no consequences associated with their test performance. As a result, to the extent that test takers respond non-effortfully, the resulting test scores will tend to underestimate what the test takers know and can do.

The reality that unmotivated test taking sometimes occurs has led researchers to consider measurement methods that can manage or accommodate its presence. Two general approaches have been proposed. The first is *individual score validation (ISV)*, in which test scores are classified as invalid if some pre-specified motivation criterion is met. These classifications can be based on a variety of criteria, such as proctor observation, test taker self-reports, person-fit statistics, or other test-taking behaviors, such as response time (S. L. Wise, in press). The intent of ISV is to identify scores that are so likely to be distorted by low motivation that they are untrustworthy indicators of test taker achievement. These scores can subsequently be flagged on score reports, or potentially deleted from psychometric analyses that use aggregated test data.

In the second approach, a test event is conceptualized as a series of items administered to a test taker, and the test taker's motivation may vary across items. Based on this conceptualization, researchers have proposed several types of psychometric models for representing how motivation changes during a test event and how it affects item responses. The

most common type is what we will call *absorbing state models* (Bolt, Cohen, & Wollack, 2002; Cao & Stokes, 2008; Jin & Wang, 2014; L. L. Wise, 1996; Yamamoto & Everson, 1995)[1]. In these models, all test takers begin in a motivated state, but during the test some switch to an unmotivated state in which they begin giving random responses. The point at which a switch occurs can vary across test takers, but once they enter the unmotivated state they remain there. A second type of model, characterized by *decreasing effort*, is similar to the absorbing state models except that instead of a sudden switch to random responding, some test takers begin exhibiting gradually decreasing effort (Cao & Stokes, 2008; Goegebeur, De Boeck, Wollack, & Cohen, 2008). The third type is a *difficulty-based model* (Cao & Stokes, 2008), in which unmotivated test takers respond randomly only when they receive items that are very difficult for them.

The three types of models characterize unmotivated test taking in different ways, and each is consistent with a characteristic pattern of responding. Ideally, one should choose a psychometric model that adequately represents how test takers actually behave. In practice, however, it is difficult to confidently conclude that a particular item response is effortful or non-effortful just by looking at its correctness. Incorrect answers may result from effortful test taking (due to lack of knowledge), while correct answers may result from non-effortful test taking (due to lucky guessing). Such uncertainty complicates interpretations of test taker behavior. Empirical support for a model's use is typically based on an examination of the model's fit statistics. If the fit statistics for the model under consideration have acceptable values and indicate better fit than those from alternative models, then use of the model is supported.

Although fit statistics are useful general tools for understanding how well a model fits a particular data set, more fine-grained information about behavior can be provided by item response time. Schnipke & Scrams (1997) studied test taker behavior on high-stakes multiple-

choice tests. They observed that, as the time limit approached, many test takers would switch strategies from trying to effortfully answer items (termed *solution behavior*) to rapidly entering answers to remaining items in hopes of getting some correct by chance (termed *rapid-guessing behavior*). Solution behavior is differentiated from rapid-guessing behavior by the time it takes a test taker to respond to an item. Schnipke and Scrams found that the accuracy rates of rapid guesses closely resemble the accuracy rates expected from random responding. Thus, in high-stakes testing contexts, rapid guessing behavior reflects a strategic choice by test takers in an attempt to maximize their scores as time is expiring.

S. L. Wise and Kong (2005) studied the data from *unspeeded, low-stakes* tests and discovered instances of rapid-guessing behavior occurring throughout test events, and not just at the end as Schnipke and Scrams (1997) had observed with speeded, high-stakes tests. Wise and Kong showed that in low-stakes contexts rapid guesses indicate unmotivated test taking. These types of rapid guesses have been found to have accuracy rates resembling those from random responding, and a number of studies have found additional evidence supporting the conclusion that rapid guessing indicates unmotivated test taking (see S. L. Wise, in press).

A rapid guessing-based approach to identifying unmotivated test taking has the important feature of allowing motivation to be evaluated down to the level of individual item responses. This suggests that an examination of rapid guessing patterns could provide valuable information about how unmotivated test taking occurs during test events. The purpose of the present investigation, which focused on data from an adaptive multiple-choice achievement test, was twofold. The first purpose was to examine patterns of rapid-guessing behavior and compare them to the patterns expected under absorbing state, decreasing effort, and difficulty-based models of test-taking behavior. The second purpose was to study the degree to which

measurement is improved by employing an item response theory (IRT) scoring model that incorporates rapid-guessing behavior. S. L. Wise and DeMars (2006) proposed their *effort-moderated IRT model*, in which two different item response functions are specified—one for solution behavior and one for rapid-guessing behavior. Under solution behavior the probability of a correct response to an item increases with a test taker's achievement level, and can be effectively modeled with a monotonically increasing function such as that represented under a traditional IRT model. In contrast, under rapid-guessing behavior the probability of a correct response to a particular item is modeled as a constant value at (or near) chance level regardless of the test taker's achievement level.

Under the effort-moderated model, each item response is modeled by one of the two response functions, depending on how quickly the test taker responds to the item. Each item response is classified as either a solution behavior or a rapid guess by comparing response time to a predetermined time threshold. Thus, for item *j*, there is a threshold T_j that differentiates rapid-guessing behavior from solution behavior. Given a test taker *i*'s response time, RT_{ij}, to item *j*, a dichotomous index of solution behavior, SB_{ij}, is computed as

$$SB_{ij} = \left\{ \begin{array}{l} 1 \text{ if } RT_{ij} \geq T_j, \\ 0 \text{ otherwise.} \end{array} \right\}. \quad (1)$$

If solution behavior is represented by the Rasch model, for example, and rapid-guessing behavior is represented by a constant probability model specified as $P_i(\theta) = 1/d_j$, where d_j is the number of distractors for item *j*, the effort-moderated model would be

$$P_i(\theta) = \left(SB_{ij}\right)\left(\frac{e^{\left(\theta - b_j\right)}}{1+e^{\left(\theta - b_j\right)}}\right) + \left(1 - SB_{ij}\right)\left(1/d_j\right). \quad (2)$$

S. L. Wise and DeMars (2006) found that, when rapid guessing was present, the effort-moderated model showed better model fit than a traditional IRT model. In addition, maximum-

likelihood estimates of achievement from the effort-moderated model were found to be more valid estimates of test taker achievement.

Under the effort-moderated model the probability of passing an item under rapid guessing is assumed to be the same regardless of the achievement level of the test taker. This implies that rapid guesses are *uninformative*, because they do not provide useful information for estimating a test taker's achievement level. As a result, scoring under the effort-moderated model is equivalent to excluding rapid guesses when calculating likelihood functions. On a 50-item test, for example, if a test taker exhibited 40 solution behaviors and 10 rapid guesses, an effort-moderated achievement estimate is based only on the 40 responses that were solution behaviors. This shows that effort-moderated achievement estimates are often based on reduced numbers of item responses, with the accompanying cost of higher standard errors than those associated with achievement estimates from test takers who exhibited solution behavior to all 50 items.

Unlike the absorbing state, decreasing effort, and difficulty-based models, the effort-moderated model makes no assumptions about patterns of motivational behavior that are exhibited during a test event. Its use is based, however, on two key assumptions. First, rapid guesses are assumed to be non-effortful. This assumption, which provides the logical basis for excluding rapid guesses from scoring, has been supported by research showing the accuracy rates of rapid guesses to be close to the accuracy rate expected under random responding. Second, the model assumes that solution behaviors represent effortful responses from the test taker. This assumption is important to the validity of the effort-moderated model, because it implies that solution behaviors provide valid information about test taker achievement, even during test events in which substantial amounts of rapid guessing occurred. One of the goals of the present investigation was to further evaluate the degree to which the second assumption is supported.

Study 1

The first study pursued two research questions. First, what can an examination of rapid-guessing behavior reveal about the patterns of motivation test takers exhibit during an adaptive achievement test? Second, to what extent is measurement improved by scoring test events using an effort-moderated IRT model?

Data Source and Method

The data for this study are based on an achievement testing program in mathematics. The testing records were drawn for 285,230 students in grades 2-12 who were administered mathematics tests during the spring testing term of the 2012-2013 academic year in a single U.S. state. All tests were part of Northwest Evaluation Association's *Measures of Academic Progress* (MAP) multiple-choice testing system, which administers computerized adaptive tests (CATs). MAP assessments are administered using liberal time limits, which results in test events that are essentially unspeeded. MAP is not considered to be a high-stakes test, because test performance does not typically count toward a student's course grade.

The CAT that administered the MAP assessment used Bayesian item selection, with the final score based on a Rasch-based maximum-likelihood achievement estimates (MLE). Achievement estimates were expressed as scale scores (*RIT*) on a common scale with a mean of 200 and a standard deviation (logit) of 10. Items were selected from a large pool containing 1383 items. Nearly all of the items (98%) had five response options, with the remainder having four options. The students in the sample had been administered the MAP assessment earlier in the school year, and the starting item difficulty value for beginning each student's test event was set slightly higher than the final *RIT* from the previous administration. For most of the test events (79%), each student was administered 50 operational items. For the remaining test events

the standard error criterion required to terminate the test after 50 operational items was not met, resulting in test events in which either 1 or 2 additional operational items were administered.

Each item response was classified as either a solution behavior or a rapid guess, using time thresholds based on the normative threshold method (S. L. Wise & Ma, 2012). In this method, the threshold for an item was set at 10 percent of the average time students have historically taken to answer the item, with a maximum threshold value of 10 seconds. For example, the threshold for an item whose average response time has been 30 seconds was set at 3 seconds, whereas the threshold for an item whose response time averaged 120 seconds was set at 10 seconds. Item responses occurring faster than the item's threshold were classified as rapid guesses, with the remaining responses classified as solution behaviors.

The classification of the item responses as solution behaviors or rapid guesses was also used to calculate each student's *response time effort* (*RTE*; S. L. Wise & Kong, 2005). *RTE*, which equals the proportion of item responses during a test event that are solution behaviors, provides an overall measure of each student's test-taking motivation. *RTE* was used to identify test events materially affected by rapid guessing. S. L. Wise (in press) suggested that the scores from test events whose *RTE* values were less than .90 would be so distorted by rapid guessing that they should be classified as invalid. For the mathematics MAP assessment, this corresponded to six or more rapid guesses.

Results and Discussion

Analyses of Rapid Guessing Patterns. Table 1 shows the percentages of solution behaviors and rapid guesses in the test data. Across all students and test events, about 1% of the item responses were rapid guesses, with a mean accuracy resembling what would be expected

from random responding[2]. The accuracy rate of solution behaviors was also close to the .50 value expected from a CAT whose items were selected using a maximum information criterion.

Table 1

Percentage and Response Accuracy of Solution Behaviors and Rapid Guesses

	Percentage of Responses		Mean Response Accuracy	
Response Group	SB	RG	SB	RG
All responses	99%	1%	.52	.21
Responses occurring after the 1st rapid guess	81%	19%	.52	.21
Responses occurring after the 6th rapid guess	67%	33%	.53	.22

Note. SB = solution behavior; RG = rapid-guessing behavior.

Of particular interest was response accuracy after a student began to exhibit rapid-guessing behavior, because it indicated that the student had begun responding in an unmotivated fashion. An absorbing state model would predict that once students began to behave non-effortfully, they would continue to do so for the remainder of their test events. As Table 1 shows, this was clearly not the case for the mathematics CAT data. After the occurrence of the first rapid guess, 81% of the remaining responses were solution behaviors; after 6 rapid guesses, 67% were solution behaviors. Moreover, once rapid guessing had begun, the mean accuracy rates of solution behaviors remained around .50, supporting the conclusion that the responses classified as solution behaviors were effortful.

Figure 1 displays item-by item information for three test events that illustrate some of the patterns of test-taking behavior that occurred. In the body of the graphs, the difficulty of each item is indicated, along with its closely matched provisional *RIT* value (to which item difficulty was matched by the item selection algorithm). In addition, the final MLE is indicated. Along

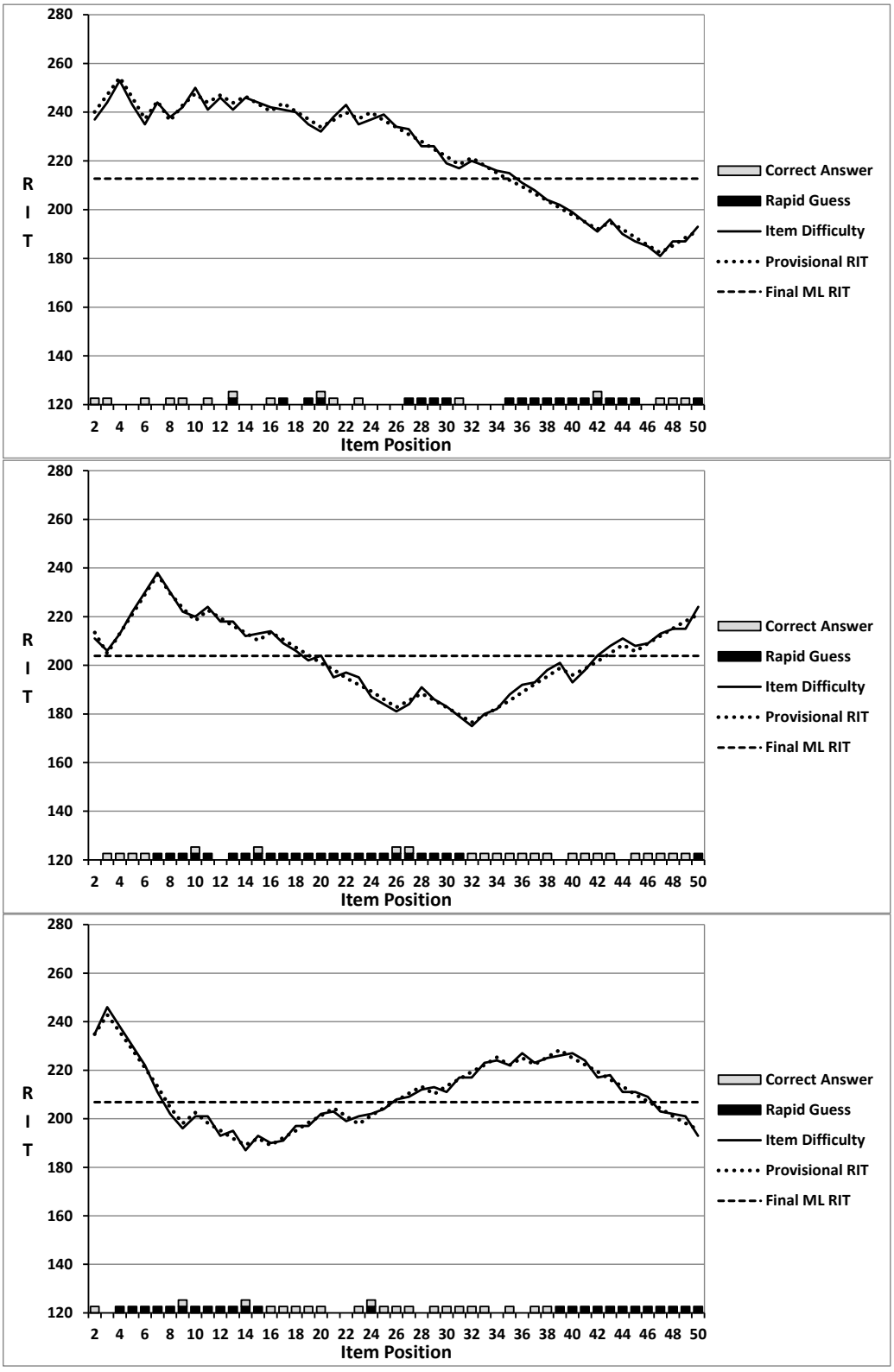

Figure 1. History graphs for three examinees taking a Math CAT. On the *RIT* scale, ten points equals one logit.

the horizontal axis the correctness of each item response is indicated, along with whether or not that response was a rapid guess.

The upper graph of Figure 1 shows a case in which the student appeared to become disengaged midway through the test event. The student's provisional *RIT* value consistently stayed near 240 until around the 25th item, after which it showed a gradual decline down to around 190—a decrease of five logits. The final MLE, which is based on all of the item responses, was around 215. If the student's actual *RIT* was 240, the MLE underestimated achievement by about 25 points (2.5 logits). The student's performance decline coincides with the student's exhibiting rapid guessing on 17 of the last 25 items. Note, however, that during the last half of the test the student passed only one rapid guess, but nearly half of the solution behaviors. Thus, although the graph is somewhat consistent with what would be expected under an absorbing state model, solution behaviors continued to occur after rapid-guessing behavior had commenced.

The middle graph in Figure 1 tells a different story. In this test event, after 5 solution behaviors, the student exhibited a nearly unbroken string of 25 rapid guesses (only 4 of which were correct). At this point, the student's behavior dramatically changed, and there was a string of 18 solution behaviors (nearly all of which were correct). If the student had passed into an unmotivated state beginning at item 6, they appeared to *re-engag*e at item 32—a finding that is clearly inconsistent with absorbing state models. From the perspective of adaptive testing, this test-taking behavior raises an additional concern. If this student's actual achievement level lay near where they began and ended the test (around 220), when re-engagement occurred at item 32, the item selection algorithm administered an item whose difficulty was mis-targeted by roughly 4 logits. This item provided very little IRT information, and it took a number of additional items before the mis-targeting was reduced to the point that subsequent items began to

provide the desired amount of information. This illustrates that unmotivated test taking can distort the difficulty targeting feature of a CAT, and thereby diminish its efficiency.

The lower graph in Figure 1 shows multiple behavior changes. In this test event, the student appears unmotivated during the initial 15 items, then motivated until item 39, and then unmotivated for the remainder of the items. Correspondingly, the accuracy rates of the item responses are low during the unmotivated sections and high during the motivated sections.

Collectively, Table 1 and Figure 1 indicate that an absorbing state model poorly describes the rapid-guessing behavior found in the study data. Although rapid guesses often appeared in clusters during test events, consistent patterns for their occurrence were not observed. Moreover, as Table 1 indicated, after rapid guessing had begun to be exhibited, solution behaviors remained far more likely than rapid guesses to be exhibited in subsequent item responses.

Rapid-guessing behavior also provided insight regarding decreasing effort models, which posit that some test takers begin to exhibit gradually decreasing effort throughout the remainder of their test events. To evaluate this type of model, we focused our attention on test events from the unmotivated 5,567 students whose *RTE* value was less than .90. Figure 2 shows that the percentage of rapid guesses occurring at each item position showed a gradual increase from less than 5% at the beginning of the test to roughly 35% near the end. While these percentages appear to be consistent with a conceptualization of gradually decreasing effort, the accuracy rates of rapid guesses and solution behaviors suggest a somewhat more complex explanation. To understand this, it is important to note that at the beginning of a MAP test event, the beginning items are purposefully selected to be relatively easy for the student. This allows the student to "settle in" to the test event. After these initial items, subsequent items have roughly a .50 probability of being passed (based on the maximum information item selection). Figure 2 shows

that, after the initial items, the proportion of correct solution behaviors did not tend to decrease during subsequent items. In fact, it tended to modestly increase due, at least in part, to the fact that unmotivated students tended to receive items that were mis-targeted on the easier side. The accuracy rates for rapid guesses were consistent across item position and the accuracy rate for all responses decreased until the 20th item position, and then remained relatively consistent throughout the remaining positions. Collectively, Figure 2 indicates that while the percentage of rapid guesses increased during the test event, the accuracy of solution behaviors did not decrease. That is, the solution behaviors did not reflect decreasing effort, but the rapid guesses reflected decreasing effort simply because there were more of them. This bifurcated view appears to more accurately characterize test taking behavior than that posited by decreasing effort models, for which there would be generically a gradually diminishing probability of passing an item.

The difficulty-based model of test-taking motivation assumes that an unmotivated student would respond randomly to items that are too difficult. This suggests that for these students there should be a systematic relationship between the difficulty levels of items administered to a student and occurrences of rapid guessing behavior. However, whether a particular item is "too difficult" depends on the achievement level of the student. To take this into account, correlations between item difficulty and rapid-guessing behavior (0 = solution behavior; 1 = rapid guessing) were calculated for the test events of the students whose *RTE* was less than .90. The difficulty-based model would predict that these correlations should be positive. The distribution of correlation coefficients, however, was found to be approximately normal with a mean of -0.05 and a standard deviation of 0.24. Thus, rapid-guessing behavior appeared to be unrelated to item difficulty, which is inconsistent with the difficulty-based model of motivation.

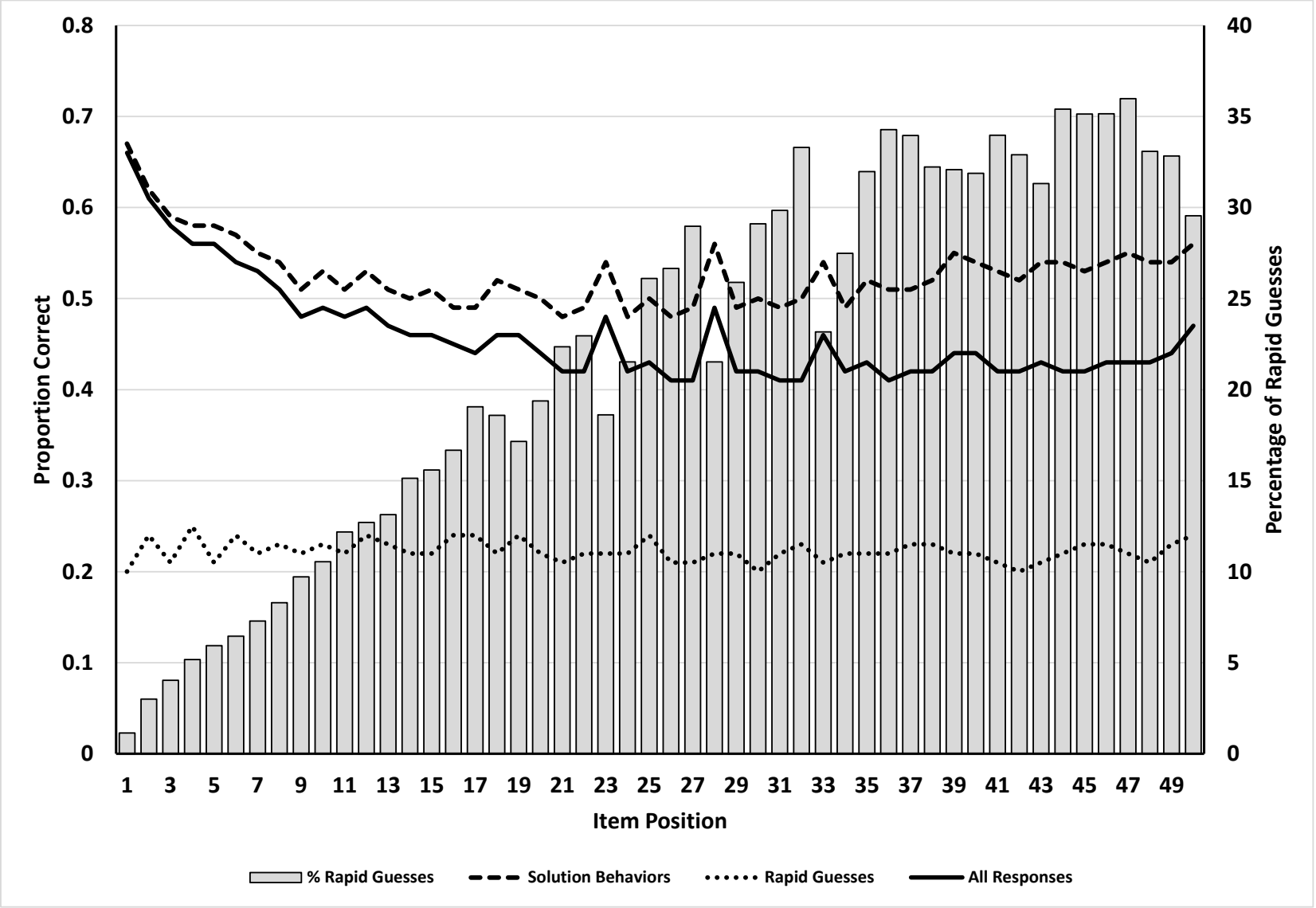

Figure 2. Accuracy of solution and rapid-guessing behaviors, by item position, for students with *RTE* less than .90 (N = 5,567).

To summarize, our analyses of rapid-guessing behavior found little evidence to support either absorbing state or difficulty-based models of motivation. Although we found that rapid guessing tended to increase *on average* across item positions, unmotivated test taking appeared to be better characterized as increasing instances of random responding to some items rather than a general decrease in effort to all items. The finding that responses classified as solution behaviors continued to be passed at a rate near .50 even after multiple rapid guesses had been exhibited supports this bifurcated conceptualization over that of a decreasing effort model.

Overall, our analyses suggest that rapid-guessing behavior on MAP is best characterized as idiosyncratic. This does not mean that rapid guessing is without systematic influences, as multiple correlates of rapid guessing have been found (S. L. Wise, 2006; S. L. Wise, Pastor, & Kong, 2009). Moreover, S. L. Wise and Smith's (2009) model of test-taking motivation posits that whether or not a test taker responds effortfully to an item is influenced by three types of factors: characteristics of the item, characteristics of the test taker, and the context in which the item is administered. The potential presence of multiple potential influences on test-taking motivation may render it difficult to predict when rapid guessing will occur.

Analysis of the Effort-Moderated Model. An important characteristic of the effort-moderated model (see Equation 2) is that it assumes no pattern of test-taking motivation across items. Instead, it focuses on classifying each item response as either solution behavior of rapid-guessing behavior. Figure 2 supports its assumption that rapid guesses are non-effortful. In addition, the effort-moderated model also specifies that the accuracy of rapid guesses should be constant across the range of student achievement. This was observed in the MAP data; to illustrate this, we divided the 9th graders in the sample into quintiles based on their achievement estimate at the beginning of their test event (which was undistorted by rapid guessing). The

accuracy rates of subsequent rapid guesses across the quintiles were .23, .22, .22, .22, and .23, respectively. This supports the effort-moderated model's assumption of rapid guessing having constant accuracy rates across achievement level, and underscores the basic principle of the model that rapid guesses are uninformative about a student's achievement level.

The second assumption—that responses classified as solution behaviors were effortful—was also supported by the data. Both Table 1 and Figure 2 show that solution behaviors exhibited accuracy rates very close to the .50 value expected under the CAT item selection algorithm. Moreover, the accuracy rates remained far above chance level even for students who had exhibited material amounts of rapid guessing.

The impact of effort-moderated model on achievement estimation is shown in Figure 3. The difference between the *RIT* values under effort-moderated scoring and traditional MLE scoring tends to increase as values of *RTE* decrease. In the most extreme cases of rapid guessing (i.e., *RTE* values around .20), this difference reached as high as six logits. If one were to assume that the scores based on the effort-moderated model are correct, Figure 3 indicates the bias in MLE scores that were due to the distortive effects of rapid guessing. At any value of *RTE*, the vertical differences in the data points are due to the accuracy rate of the rapid guesses. If the student was a relatively unlucky guesser, bias was relatively high; whereas, if the student was a lucky guesser, bias was lower. In a small proportion of cases, the effort-moderated score was actually lower than the MLE score. This occurred when the student's rapid guessing accuracy rate exceeded their solution behavior accuracy rate.

The effect of effort-moderated scoring on the standard errors of the achievement estimates is shown in Figure 4, which illustrates the loss in information associated with the effort-moderated model. As *RTE* decreases, the number of informative item responses used in

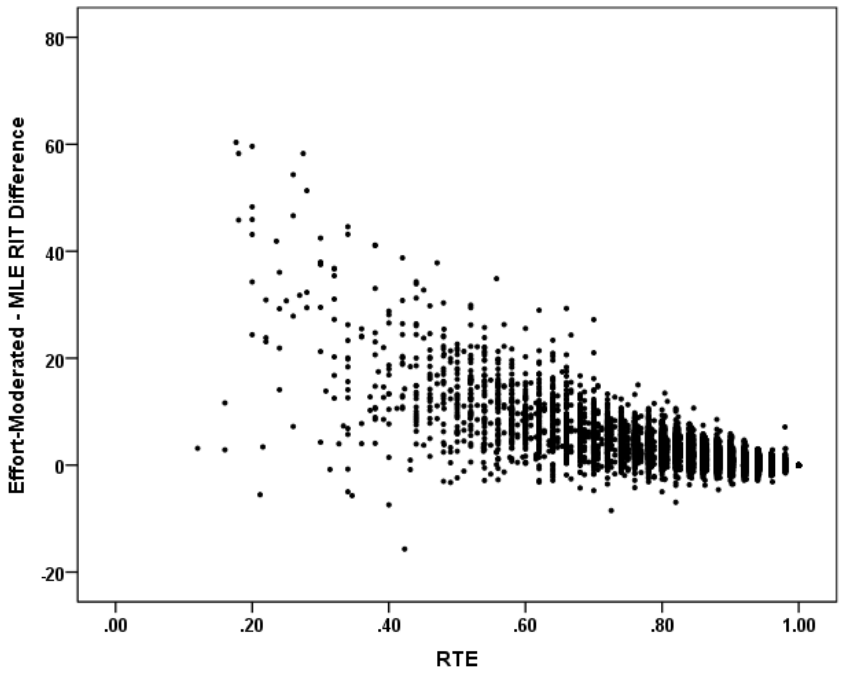

Figure 3. Scatterplot of the relationship between *RTE* and the difference in *RIT* scores between MLE and effort-moderated scoring. On the *RIT* scale, ten points equals one logit.

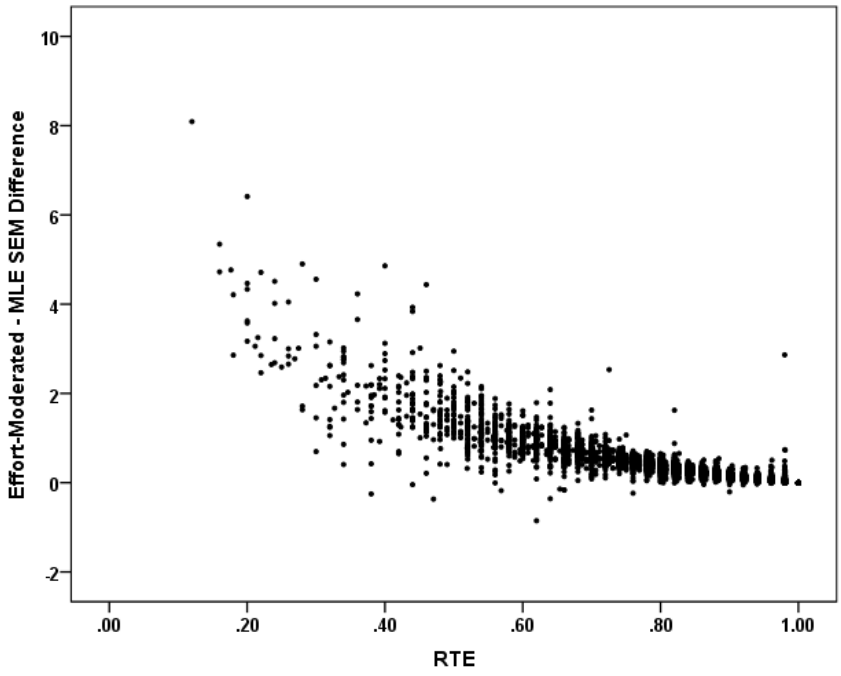

Figure 4. Scatterplot of the relationship between *RTE* and the difference in standard errors between MLE and effort-moderated scoring. The standard error of Math *RIT* scores is typically around three points.

estimating achievement correspondingly decreased. This resulted in effort-moderated achievement estimates with larger standard errors than those found under MLE scoring.

Our interpretation of the score differences shown in Figure 3 presumes that the effort-moderated model provided more accurate scores, and that the traditional MLE scores were biased. This assertion was assessed using a likelihood-based person fit statistic l_z (Drasgow, Levine, & Williams, 1985). For each of the test events in the sample, l_z was computed based either on all of the responses or only for those responses that were solution behaviors (i.e., those used by the effort-moderated model). The theoretical sampling distribution of l_z has been shown to approximate a standard normal distribution, with values of zero indicating perfect model fit, and negative l_z values indicating that the student had performed more poorly than would be expected by the IRT model. The median values of l_z for each model are shown in Figure 5.

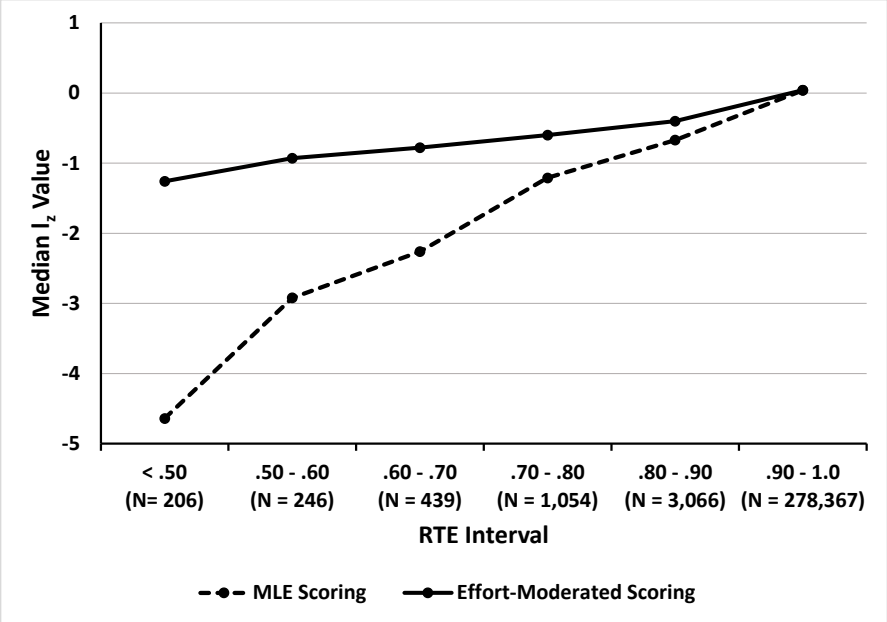

Figure 5. Impact of effort-moderated scoring on median person fit indices (l_z), by *RTE* interval.

When *RTE* was at or near 1.0, person fit was equivalent for each model, with median l_z values near 0.0. As *RTE* decreased, however, median person fit decreased sharply under MLE scoring,

while values for the effort-moderated model showed a gradual decrease. Thus, in the presence of rapid guessing, effort-moderated scoring exhibited superior person fit, supporting the interpretation that its achievement estimates were less biased than those from MLE scoring.

Overall, the results for the effort-moderated model showed that its assumptions were satisfied by the MAP data, and person fit statistics indicated that the model fit the student data better than did the traditional IRT model. Furthermore, it was found that the amount of bias in achievement estimates due to rapid guessing could be as high as six logits in magnitude, and that the effort-moderated model could effectively reduce the bias.

Study 2

The second study was intended to study two additional research questions. The first concerned the degree to which the effort-moderated model corrects for the score distortion due to unmotivated test taking. Bowe, Wise, and Kingsbury (2011) investigated the effort-moderated model using MAP data, noting that in test events in which marked amounts of rapid guessing were present, it was difficult to know the extent to which solution behaviors might also be influenced by unmotivated test taking. They suggested that scores from the effort-moderated model may still retain some negative bias. Although the results of our Study 1 suggest that solution behaviors appeared to be generically effortful—regardless of the presence or absence of rapid guessing—Study 2 sought to further clarify the matter.

The second research question was based on our finding in Study 1 that some students appeared to re-engage in their test following a period of substantial rapid guessing. Once re-engagement occurred, many of the subsequent items provided little information (in an IRT sense) for estimating achievement. A possible solution to this problem would be for the CAT algorithm to be modified such that it would monitor test-taking behavior during the test event and not alter

its provisional *RIT* value after a rapid guess had occurred. This *effort-guided* modification was expected to prevent the type of mis-targeting observed in the Study 1 data, and thereby yield scores with smaller standard errors. Therefore, the second research question concerned the impact of an effort-guided CAT on the standard errors of achievement estimation.

Method and Data Generation

To address the two research questions, three types of CATs were independently simulated as follows:

- The first simulated adaptive test was a traditional CAT that used Bayesian provisional *RIT* values used during item selection, with a final MLE achievement estimate (MLE-CAT). Randomesque item selection was used to randomly select each item from the 15 most informative items at the provisional achievement level estimate, and each test had a fixed length of 50 items. This test served as a baseline, since it ignored test taker motivation in item selection and scoring.
- The second simulated adaptive test was identical to the MLE-CAT, but the final score was based on the effort-moderated model (EM-CAT). This test still ignored test taker motivation in item selection, but the scoring algorithm excluded rapid guesses in calculating a test taker's final score.
- The third simulated adaptive test was an effort-guided CAT (EG-CAT), for which (a) the provisional achievement estimate remained unchanged after each rapid guess and (b) effort-moderated scoring was used. This test was designed to test taker motivation into account during both item selection and scoring.

Item Data. The three simulated CATs used an item pool whose difficulty parameters were identical to those in the real mathematics item pool used in Study 1.

Test Taker Data. Each of the simulated tests used a sample of 51,764 test takers. Half of the sample consisted of motivated test takers who exhibited no rapid guessing, while the other half comprised unmotivated test takers who exhibited varying amounts of rapid-guessing behavior. To realistically simulate unmotivated test taking, the patterns of solution behaviors and rapid guesses exhibited by students in Study 1 were used as motivational "templates" that were applied to the behavior of simulated test takers in Study 2. For example, if there was a student in Study 1 who had given 12 rapid guesses during a test event, there was a corresponding test taker for each Study 2 CAT that exhibited 12 rapid guesses in the exact same item positions during their simulated test events. This assured that (a) each of the three tests had test events with rapid guessing patterns that had actually been observed in real data, and (b) the CATs were based on equivalent numbers of rapid guesses. For each simulated CAT test event, the true achievement level was set equal to the achievement level estimate used to select the first item in the corresponding motivational template test event from Study 1.

Results and Discussion

Table 2 compares the three types of CATs in terms of bias and standard errors. The MLE-CAT showed mean bias that was inversely related to *RTE*. The bias could be sizeable; for test takers with *RTE* less than .50, mean bias exceeded 16 *RIT* points. In contrast, both the EM-CAT and the EG-CAT exhibited negligible mean bias throughout the *RTE* range. The CATs also differed in the standard errors of achievement estimation, as shown in Table 2. The MLE-CAT, whose scores were always based on the full set of item responses, consistently showed smaller mean standard errors than did the EM-CAT and the EG-CAT, with the differences increasing as *RTE* decreased. These results show that the CAT types using effort-moderated scoring yielded scores with less bias, but also with less precision than the MLE-CAT.

Table 2

Bias, SEM, and Recovery Rates of True Achievement Level for the Three Simulated CATs

Type of CAT	*RTE* Interval					
	< .50	.50 - .60	.60 - .70	.70 - .80	.80 - .90	.90 - 1.0
			Mean Bias			
MLE-CAT	16.11	9.19	6.34	3.85	2.01	0.19
EM-CAT	-0.13	-0.03	-0.07	-0.10	-0.03	-0.02
EG-CAT	0.02	-0.02	-0.07	-0.10	0.00	-0.01
			Mean SEM			
MLE-CAT	3.11	3.01	2.98	2.96	2.94	2.96
EM-CAT	5.37	4.10	3.73	3.43	3.21	3.00
EG-CAT	4.90	3.98	3.65	3.38	3.18	2.98
			Percentage of 95% Confidence Intervals Capturing the True Achievement Level			
MLE-CAT	16	33	50	70	85	94
EM-CAT	95	95	95	94	95	95
EG-CAT	95	95	95	95	95	95

Note. Bias and SEM are expressed on the *RIT* scale, for which 10 *RIT* points equals 1 logit.

To assess whether the effort-moderated model's loss in precision was worth the decrease in bias, 95% confidence intervals were constructed for each test event. Table 2 shows the percentage of confidence intervals containing the true achievement level for each test type. For the MLE-CAT, the percentages increasingly fell below 95% as *RTE* decreased, while the percentages remained at the nominal 95% level throughout the *RTE* range for both EM-CAT and EG-CAT. This showed that, in the presence of rapid guessing, scores from the CATs using effort-moderated tended to be less precise but more accurate than those based on MLE scoring.

Table 2 also shows that the standard errors of scores from the EG-CAT tended to be lower than those from the EM-CAT. This difference reflects the vulnerability of the EM-CAT to mis-targeted item difficulty for unmotivated students who re-engaged during a test event.

Although the mean standard error tended to be only modestly lower for the EG-CAT, there were several instances in which the difference was large. In particular, under the EM-CAT condition three cases with *RTE* less than .20 were found to exhibit very large standard errors of 37, 40, and 56 *RIT* points, respectively—indicating severe mis-targeting. For these same cases, under the EG-CAT the respective standard errors were 11, 13, and 14 *RIT* points. Hence, the most extreme cases of mis-targeting could be substantially ameliorated using the EG-CAT.

The simulated data from Study 2 additionally provided a useful reference point for evaluating the degree to which the effort-moderated model's assumption of effortful solution behaviors was met by the empirical data in Study 1. Essentially, Study 2 simulated a scenario in which the assumptions of the effort-moderated model were fully satisfied: solution behaviors conformed to the IRT model, whereas rapid guesses conformed to a flat, constant-probability function. The relationship between *RTE* and the *RIT* differences between the simulated MLE-CAT and EM-CAT conditions is shown by the dotted line in Figure 6. This non-linear regression line depicts the accelerating score estimation bias associated with decreasing *RTE*. The solid line in Figure 6 shows the corresponding regression line for the empirical data. To the extent that unmotivated test taking reduced the probability of passing items under solution behavior, the regression line for the empirical data would be expected to lie below that for the simulated data. The figure shows, however, that the two regression lines were highly similar, with the empirical line lying slightly above that from the simulated data. This implies that the solution behaviors in the empirical data were relatively unaffected by unmotivated test taking.

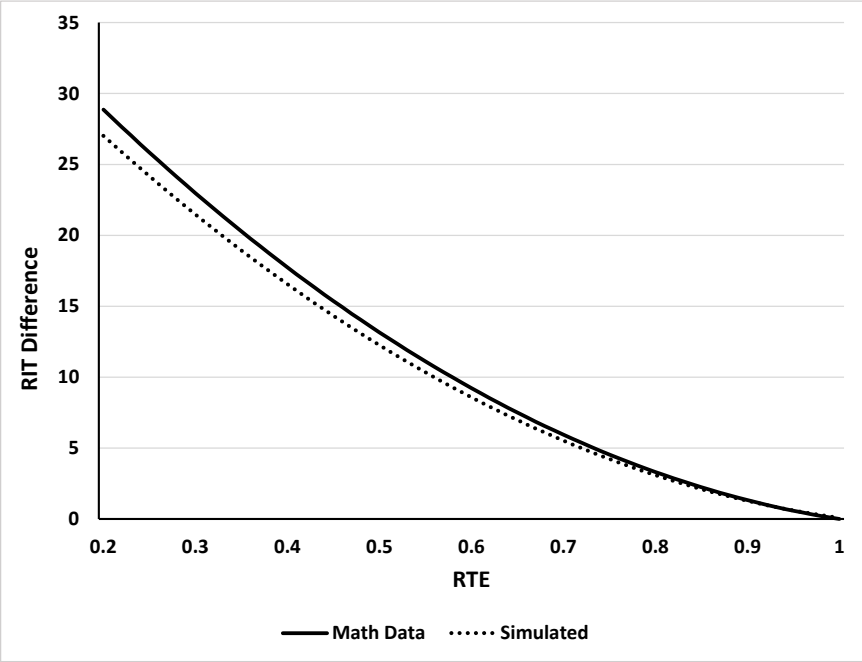

Figure 6. Comparison of non-linear (linear + quadratic) regression lines from the data in Studies 1 and 2. The solid line shows, for the actual data, the relationship between *RTE* and the *RIT* difference between MLE-CAT scoring and effort-moderated scoring. The dotted line shows, for the simulated data, the relationship between *RTE* and the *RIT* difference between MLE-CAT and EM-CAT scores.

General Discussion

Because our traditional measurement models typically do not acknowledge the influence unmotivated test taking, concerns for validity encourage us to develop innovative measurement models and methods that can accommodate its presence. This endeavor is hindered by the fact that there is much that we still do not know about the behavior of unmotivated test takers during a test event. Diffuse types of indicators such as relative model fit statistics are limited in the specific information they can provide about how test takers behave. One purpose of this paper was to demonstrate that analyses of rapid-guessing behavior can provide useful valuable item-by-item information about test-taking motivation. We found that for the MAP data studied in this investigation, the rapid-guessing behavior exhibited by students was inconsistent with what

would be expected by absorbing state or difficulty-based models. In addition, although there was a clear indication of increased rapid guessing as item position increased, the accuracy rates of solution behaviors did not decrease, which was contrary to what would be expected under a decreasing effort model.

The results of our studies of MAP data suggest a bifurcated view of test-taking motivation. Rapid guesses appeared to occur idiosyncratically, and were correct at a rate consistent with random responding. Solution behaviors appeared to be correct at a rate consistent with that expected during adaptive testing, regardless of whether or rapid guessing was occurring during a test event. It should be emphasized, however, that one should be cautious in generalizing the conclusions from this study. Other tests, administered in other measurement contexts, may yield patterns of rapid guessing and solution behaviors that are more consistent with absorbing state, decreasing effort, or difficulty-based models. We do suggest that a routine analysis of rapid-guessing behavior could help measurement practitioners better understand how best to model unmotivated test taking in their testing program.

Another purpose of this investigation was to evaluate the usefulness of the effort-moderated model (S. L. Wise & DeMars, 2006). When rapid guessing is present during a test event, effort-moderated scoring was found to both yield scores that were more accurate than traditional MLE scores, and exhibited improved person fit statistics. In addition, the effort-guided modification of the effort-moderated CAT was found to yield more precise scores by effectively controlling the item difficulty mis-targeting problem experienced by test takers who re-engage after a period of disengagement.

The effort-moderated model has several advantages that make it desirable for practical use when computer-based testing is used. The first is its simplicity; it does not require

estimation of additional model parameters, as is typical of absorbing state and decreasing effort models. It merely identifies rapid guesses and excludes them from achievement estimation. The second advantage is its testability; its assumptions of effortful solution behaviors and non-effortful (and uninformative) rapid guesses can be readily evaluated using the methods described in this paper. The third advantage is its flexibility; it does not require the user to make assumptions about patterns of rapid guessing that test takers exhibit. It could be applied when test takers behave in accordance with either absorbing state, difficulty-based, or many other potential models. It could also be applied in situations in which some test takers' behavior is consistent with one model, while the behavior of other test takers conforms to a different model.

It should be noted that there are practical limitations on the application of the effort-moderated model. For example, although one *could* apply effort-moderated scoring to a test event in which only 20% of the responses were solution behaviors, it might not be desirable to do so. Because the standard error of such a score would probably be unacceptably large and desired item content balance for a test event may not have been attainable, the credibility of scores based on so few item responses would be diminished. In practice, a policy would need to be adopted regarding the minimum number of solution behaviors (and their content coverage) that would be needed during a test event for a score to be valid.

This study has improved our understanding of the dynamics of test-taking motivation, and has shown the value of analyses of rapid-guessing behavior. The item-by-item information about motivation provided by rapid guessing is incorporated in the effort-moderated model, which can improve the quality of measurement and thereby improve the validity of score-based inferences.

References

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement, 40*, 109-128.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*, 331-348.

Bowe, B., Wise, S. L., & Kingsbury, G. G. (2011, April). *The utility of the effort-moderated IRT model in reducing negative score bias associated with unmotivated examinees*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*, 209-230.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, *73*, 65-87.

Jin, K., & Wang, W. (2014). Item response theory models for performance decline during testing. *Journal of Educational Measurement, 51*, 178-200.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*, 213-232.

Wise, L. L. (1996, April). *A persistence model of motivation and test performance.* Paper presented at the annual meeting of the American Educational Research Association, Vancouver, Canada.

Wise, S. L. (in press). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education.*

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43,* 19-38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 163-183.

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, and C. W. Buckendal (Eds.), *High-stakes testing in education: Science and practice in K-12 settings* (pp. 139-153). Washington, DC: American Psychological Association.

Yamamoto, K., & Everson, H. T. (1995). *Modeling the mixture of IRT and pattern responses by a modified HYBRID model* (ETS Research Report RR-95-16). Princeton, NJ: Educational Testing Service (ERIC Document Reproduction Service No. ED 395036).

Footnotes

[1] Psychometric models under this conceptualization of unmotivated test taking have taken several forms, including Markov processes (L. L. Wise, 1996), latent class models (Bolt et al., 2002; Jin & Wang, 2014), and item response theory models (Cao & Stokes, 2008; Yamamoto & Everson, 1995).

[2] Actually, we expected rapid guessing accuracy to be slightly higher than .20, for two reasons. First, although the vast majority of the items had five response options, a small percentage had four options. Second, it has been found that test takers do not actually guess randomly, but instead tend to choose middle options more frequently (Attali & Bar-Hillel, 2003). That tendency, coupled with the fact that the MAP item pool was somewhat imbalanced regarding the position of the correct answer (with slightly higher percentages occurring in the second and third positions), could explain the slightly higher rapid guessing accuracy.