Running Head: EFFORT-MODERATED IRT MODEL

An Application of Item Response Time: The Effort-Moderated IRT Model

Steven L. Wise, Christine E. DeMars, and Xiaojing Kong

James Madison University

Paper presented at the annual meeting of the American Educational Research Association, Montreal, April, 2005.

Correspondence regarding this paper should be addressed to Steven L. Wise, Center for Assessment and Research Studies, James Madison University, MSC 6806, Harrisonburg, VA 22807. E-mail: wisesl@jmu.edu.

An Application of Item Response Time: The Effort-Moderated IRT Model

The validity of inferences based on achievement test scores is dependent on the amount of effort that examinees put forth while taking the test. With low-stakes tests, for which this problem is particularly prevalent, there is a consequent need for psychometric models that can take into account different levels of examinee effort. This article introduces the effort-moderated IRT model, which incorporates item response time into proficiency estimation and item parameter estimation. In two studies of the effort-moderated model when rapid guessing (i.e., reflecting low examinee effort) was present, one based on real data and the other on simulated data, the effort-moderated model performed better than the standard 3PL model. Specifically, it was found that the effort-moderated model (a) showed better model fit, (b) yielded more accurate item parameter estimates, (c) more accurately estimated test information, and (d) yielded proficiency estimates with higher convergent validity.

Whenever we administer an achievement test to an examinee, we tacitly assume that the examinee will try to show us what he or she knows and can do. That is, the validity of any inference made about the examinee on the basis of his or her test score is dependent on the amount of effort that the examinee put forth while taking the test. If adequate effort was not given, then performance was likely to suffer—thus resulting in the test score under-estimating the examinee's true level of proficiency.

Measurement professionals are generally aware, however, that in practice there are instances when some examinees do not give good effort to their tests. The impact of low examinee motivation can be sizable in magnitude. Wise and DeMars (in press) recently synthesized a number of studies of the effects of examinee motivation on test performance, finding that motivated examinees tend to outperform their less motivated peers by an average of 0.58 standard deviations. Wise and DeMars also reported evidence that the effort examinees expend toward their tests tends to be unrelated to ability, which suggests that the test performance differences they reported are due primarily to differences in effort, not ability.

Wise and Kong (in press) noted that there are at least three situations in which examinee effort should be a concern to measurement professionals. First, there are a number of assessment programs that have serious potential consequences for institutions but little, if any, personal consequences for individual examinees. For example, there has been a growing emphasis at the K-12 level on frequent assessment testing as a means to hold schools accountable for the quality of education that they provide their students. In particular, the recent No Child Left Behind (NCLB) legislation has profoundly expanded the presence of achievement testing in K-12 institutions. Similarly, in higher education, assessment testing is increasingly being used to hold publicly funded institutions accountable for expenditures of taxpayer dollars. Note that in each of these examples, the testing is high-stakes for the institution, but relatively low-stakes for the examinee. Throughout this article, we will characterize such testing as low-stakes, to emphasize the examinee's perspective.

Second, high-stakes testing programs sometimes administer test items in low-stakes settings, particularly in the early stages of the program. It is not uncommon, for example, for a new licensure testing program to pilot test items in non-consequential settings to obtain the psychometric data that are subsequently used to calibrate items, construct test forms, and perform additional linking/equating tasks. In these testing situations, there are typically no consequences associated with test performance for examinees.

Third, a substantial amount of measurement research is conducted in low-stakes settings at colleges and universities. This type of research frequently uses volunteers or students from subject pools, for whom there are typically minimal consequences associated with test performance. That is, these examinees may be penalized for failing to show up to participate in a study, but those who do show up will usually not be penalized for not trying very hard.

Theoretical Perspective

Any discussion of the dynamics of examinee effort and how it relates to test performance should consider contemporary models of achievement motivation. One useful theoretical basis for representing examinee behavior during test taking is provided by expectancy-value models of achievement motivation. These models specify that a student's achievement behavior when approaching a task depends on two factors: *expectancy*, which are the student's beliefs or judgments that he or she can successfully complete a task, and *value*, which are the beliefs that the student holds regarding why he or she should complete the task. Pintrich and Schunk (2002) provide a good historical and comparative overview of expectancy-value models.

An expectancy-value model of achievement motivation that is of particular relevance in a test-taking context comes from the work of Eccles and Wigfield (e.g., Eccles, 1983; Wigfield & Eccles, 2000). In the Eccles-Wigfield model, students' expectancy beliefs regarding success on a task are influenced both by their beliefs regarding their competence and by their perceptions of the difficulty of the task. In contrast, students' value beliefs are influenced by attainment value (e.g., the importance of doing well on the task), intrinsic value (e.g., the enjoyment gained from doing the task), utility value (e.g., how the task fits into one's future plans), and perceived costs (e.g., what one has given up to do the task).

In a low-stakes assessment test, there are typically few (if any) consequences associated with examinee performance and some examinees are likely to perceive minimal personal benefit from the assessment testing experience. Hence, it should be expected that these examinees would tend to hold weak value beliefs, which would lead to low examinee effort on the assessment test. For these examinees, the task of doing well on the test will have limited attainment, intrinsic, or utility value. In addition, these examinees are likely to be aware of the costs associated with the assessment test (i.e., being denied the opportunity to engage in activities that they value more highly). The Eccles-Wigfield model would therefore predict low effort (and consequently, diminished

performance) on low-stakes assessments tests from those examinees who hold weak value beliefs.

An alternative theoretical model that is relevant to assessment testing is Snow's concept of aptitude (Snow, 1989, 1992, 1994). In Snow's model, there are two parallel pathways that describe psychological-level contributions to task performance: a performance and a commitment pathway. The performance pathway describes a process by which examinees draw on relevant cognitive resources in the service of accomplishing a particular task. In contrast, the commitment process describes a process by which examinees draw on relevant conative and affective resources in the service of guiding, energizing, and regulating behavior toward accomplishing the task (Lau & Roeser, 2002). Thus, when applied to test taking, Snow's model specifies that an examinee's performance on a test will be a function of both the levels of knowledge and skills that the examinee possesses, and the commitment that he or she has to do well on the test. Without adequate commitment, test performance will suffer.

Incorporating Effort into the Measurement Process

Assuming that examinee effort is an important component of test performance, how can it be incorporated into the measurement process? An initial challenge is to obtain a valid measure of effort. There are several options available to measurement practitioners.

A commonly used method of assessing examinee effort is to administer a post-test self-report scale. Such scales typically contain a small number of Likert-type items that require little time to administer. A drawback to self-report measures, however, is that it is difficult to tell how truthful examinees will be when asked how hard they tried on an achievement test they have just taken. Some examinees who did not give good effort might report trying hard on the test because they fear disapproval or punishment from the test giver. Alternatively, because of a predisposition to attribute failure on a task to lack of effort over lack of ability (Pintrich & Schunk, 2002), some examinees who believe they did not do well on the achievement test might under-report their effort. Thus, self-report scales are vulnerable to bias through motivational processes, and it is difficult to ascertain the degree to which this is a problem with a particular sample of examinees.

Self-report information can be used to identify and remove from the sample the data from examinees reporting low effort. This procedure, termed *motivation filtering* (Wise & DeMars, in press; Sundre & Wise, 2003), has been found to improve the convergent validity of the remaining test scores.

Person-fit statistics have been proposed as an alternative to self-report scales for assessing examinee effort. These statistics are designed to identify the aberrance of examinee response patterns, typically by assessing the fit of an examinee's item response pattern to some theoretical measurement model. The type of aberrant response pattern that is most congruent with a lack of test-taking effort is random responding (Meijer, 2003). Person-fit statistics have the attractive feature of being based on observed

behavior rather than self reports from examinees, and are therefore not vulnerable to examinee presentation biases. However, they have a serious limitation; person-fit statistics have been purported to detect several types of aberrant response patterns, many of which are inconsistent with lack of student effort. For example, in addition to detecting random response patterns, it has been proposed that person-fit statistics can also detect (a) cheating, creative responding, careless responding, and lucky guessing (Meijer, 1996), (b) examinee misconceptions in cognitive diagnosis (Tatsuoka, 1996), and (c) curricular differences among schools (Harnisch & Linn, 1981). Because person-fit statistics can detect a multitude of aberrant response patterns, with numerous interpretations, it is therefore difficult to unambiguously conclude that a particular instance of misfit is due to lack of effort.

Recently, a third method has emerged for measuring examinee effort. Wise and Kong (in press) introduced a new measure of effort that was based on item response times in computer-based tests (CBTs). The rationale for this measure is based on the research reported by Schnipke and Scrams (Schnipke, 1995, 1996, 1999; Schnipke & Scrams, 1997, 2002) in their studies of the rapid responses that examinees often give during speeded, high-stakes tests. They identified two types of examinee behaviors: *solution behavior*, in which examinees actively seek to determine the correct answer to test items, and *rapid-guessing behavior*, in which they rapidly respond to items in a generally random fashion. Schnipke and Scrams observed that rapid-guessing behavior tends to occur near the end of speeded tests, when examinees rapidly respond to remaining items as time is expiring.

Wise and Kong (in press) discovered that rapid-guessing behaviors can also be found in the data from unspeeded low-stakes CBTs. Moreover, they found that these behaviors occurred throughout the test, and not just toward the end as Schnipke and Scrams (2002) had observed with speeded high-stakes tests. Figure 1 shows the frequency distribution of examinee response times for one of the items studied by Wise and Kong. It is bimodal, exhibiting the characteristic frequency "spike" for very short response times (in this case, less than five seconds) that Schnipke and Scrams observed in their studies of speeded tests.

Wise and Kong hypothesized that rapid-guessing behaviors on low-stakes tests reflected a lack of examinee effort. To evaluate this hypothesis, they developed an index, termed *response time effort (RTE)*, for measuring an examinee's overall test-taking effort.

RTE scores are based on the conceptualization that a test session is comprised of a series of examinee-item encounters. In each encounter, the examinee makes a choice to engage in either solution or rapid-guessing behavior. This choice is reflected by the time the examinee takes to respond to the item. Thus, for item i, there is a threshold, T_i, that represents the response time boundary between rapid-guessing behavior and solution behavior. Given an examinee j's response time, RT_{ij}, to item i, a dichotomous index of item solution behavior, SB_{ij}, is computed as

$$SB_{ij} = \begin{cases} 1 & if \;\; RT_{ij} \geq T_i, \\ 0 & otherwise. \end{cases} \qquad (1)$$

The index of overall response time effort for examinee *j* to the test is given by

$$RTE_j = \frac{\sum_{i=1}^{k} SB_{ij}}{k}, \qquad (2)$$

where *k* = the number of items in the test.

RTE scores represent the proportion of test items for which the examinees exhibited solution behavior. Wise and Kong (in press) investigated the characteristics of the *RTE* scores obtained from the administration of an 80-item university assessment test. They found that *RTE* scores (a) showed very high internal consistency (alpha = .97), (b) exhibited convergent validity through significant positive correlations with both self-reports of effort and person-fit statistics, (c) showed evidence of discriminant validity through near-zero correlations with Scholastic Achievement Test (SAT) scores, and (d) yielded motivation filtering effects highly similar to those found with both a self-report measure of effort administered immediately after the assessment test and a person-fit statistic calculated from the examinee response patterns. In addition, the accuracy of responses identified as reflecting rapid-guessing behavior was similar to that expected by chance (i.e., by random responding). Taken together, Wise and Kong's findings support the inference that *RTE* scores measure examinee effort.

Threshold Identification. A key issue in differentiating solution behavior from rapid-guessing behavior for each item is the identification of its time threshold, T_i. Schnipke and Scrams (1997) used a two-state mixture model to determine where the rapid-guessing and solution behavior distributions crossed for each item, which they used as the time thresholds. Schnipke (personal communication, July 6, 2004) noted, however, that the mixture model approach was not easy to implement, and that the simpler method of setting the thresholds by inspecting the response time frequency distributions and visually choosing thresholds at the end of the short time spikes works virtually as well[1]. Wise and Kong (in press) assigned to each item one of three thresholds according to the amount of reading/scanning the item required. They found that this method was fairly effective, though they needed to adjust the thresholds for 4 of their 80 items after visually inspecting the response time frequency distributions. Wise (2004) used visual inspection alone to choose thresholds for a revised, 60-item version of the CBT used by Wise and Kong (in press), finding that those thresholds worked somewhat better than those used in Wise and Kong's study.

The adequacy of the chosen time thresholds can be evaluated in two ways. First, because rapid-guessing behavior is conceptualized as an item response occurring before the examinee had a chance to read the item in an effortful manner and identify an answer, the reasonableness of an item's threshold can be evaluated by comparing it with a

combination of the amount of reading/scanning and cognitive processing required to identify the correct answer under solution behavior. Thus, for an item that asks the examinee to identify a basic fact and has a very short stem and options, a relative short (e.g., two second) threshold might be reasonable, whereas for an item that requires a lot of reading and/or asks the examinee to interpret a complex graph, a longer threshold would be reasonable (e.g., ten seconds).

A second method for evaluating the thresholds is empirical. If the thresholds have been chosen well, then it would be expected that rapid guesses should have an accuracy similar to that expected by chance (i.e., under random responding), whereas responses resulting from solution behavior should exhibit accuracy far above that expected by chance. For example, Wise (2004) reported that 25.5% of the responses under rapid-guessing behavior were correct, as compared with an expected accuracy of 25.1% for random responses. In contrast, 72.0% of the responses under solution behavior were correct. These results support the reasonableness of the chosen thresholds.

An Effort-Moderated Item Response Model. In a given examinee-item encounter, the examinee will engage in either solution behavior or rapid-guessing behavior, and the chosen response strategy will be indicated by the time it takes for the examinee to select an answer. The accuracy rates for the two strategies will typically be substantially different. Moreover, under solution behavior the probability of a correct response typically increases with examinee proficiency, and would be effectively modeled by a monotonically increasing function such as that represented under a traditional item response theory (IRT) model. In contrast, under rapid-guessing behavior the probability of a correct response remains near the level expected by chance regardless of examinee proficiency, which would be modeled by a flat item response function that specifies a constant probability of a correct response across the range of proficiency. Thus, the functional relationship between examine proficiency and the probability of a correct response will be very different under the two strategies, as illustrated by Figure 2.

We can combine the two item response functions into a single model that is moderated by response strategy. The generic model of the probability of a correct response to an item using both types of response strategies can be represented as:

$$P_i(\theta) = (SB_{ij})(\text{solution behavior model}) + (1 - SB_{ij})(\text{rapid} - \text{guessing behavior model}), \quad (3)$$

with the dichotomous SB_{ij} defined as in Equation 1. Assuming that the value of SB_{ij} is a function of examinee effort to the item, we can refer to this as an *effort-moderated item response model*. As an example, suppose that solution behavior is represented by the three-parameter logistic (3PL) IRT model, and rapid-guessing behavior is represented by a constant-probability model specified as $P_i(\theta) = 1/d_i$, where d_i is the number of response options for item i. In this case, the effort-moderated model would be:

$$P_i(\theta) = (SB_{ij})(c_i + (1 - c_i)(\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}})) + (1 - SB_{ij})(1/d_i). \quad (4)$$

In essence, because SB_{ij} takes only the values 0 or 1, the effort-moderated model specifies two item response functions—one for solution behaviors and one for rapid-guessing behaviors. Depending on the time used by examinee *j* to answer item *i*, one or the other of the response functions would be used to model the examinee's response.

The general goal of the current investigation was to develop an IRT model that incorporates examinee effort and to evaluate its usefulness when rapid-guessing behavior is present. This goal was pursued through two studies: one using real assessment data, and the other using simulated data.

Study 1

The purpose of the first study was to examine the psychometric characteristics of an effort-moderated IRT model relative to those of a standard IRT model. It represents a re-analysis of the data collected in the second study reported by Wise (2004). The two models were compared in terms of model fit, item parameter estimation, test information, and convergent validity.

Method

Examinees. The examinees in the sample were 524 mid-year sophomores who were administered a computer-based assessment test during the spring Assessment Day at a medium-sized southeastern university. On this day, classes are cancelled and all sophomores are required to be tested as part of the assessment of the General Education program. These tests are considered low stakes, as there are no personal consequences for students based on test performance. All students were required to participate in testing.

Assessment Test. The assessment test used in this study was a 60-item version of the Information Literacy Test (ILT), which is a locally-developed test used to assess student information literacy knowledge and skills. All but one of the ILT items used a multiple choice format. The numbers of response options ranged from two to five. An examinee, when administered an ILT item, had to select an answer before he or she could move on to the next item. The test administration was unspeeded; although a 60-minute time limit was imposed, all examinees completed the test within 51 minutes. Wise (2004) reported an estimated reliability of .88 for this sample of ILT scores.

During the administration of the ILT, response time was collected for each examinee-item encounter. These response times were dichotomized as SB_{ij}s using the time thresholds used by Wise (2004). If the response time was less than the threshold, then $SB_{ij} = 0$ (rapid-guessing behavior). Otherwise, $SB_{ij} = 1$ (solution behavior). *RTE* scores were then computed for each examinee. These scores had an estimated reliability of .99 (using coefficient alpha).

The ILT data set contained a modest number of unmotivated examinees. Of the 23,160 examinee-item encounters, 1,332 (5.8%) were classified as rapid-guessing

behaviors. Moreover, the results from a self-report motivation scale administered immediately after the ILT indicated that roughly 5% of the examinees reported giving little effort to the test. All of the ILT items received rapid-guessing behavior from at least one examinee, and 31% of the examinees exhibited rapid-guessing behavior on at least one item.

Item Calibration. ILT item parameter estimation for both the effort-moderated and standard 3PL model was performed using BILOG-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 2003). Though the effort-moderated model is not an explicit model available in BILOG-MG, the *a*, *b*, and *c* parameters of the effort-moderated model are equivalent to those from the standard 3PL model with rapid-guessing behaviors coded as "not administered." Because the item response function for rapid-guessing behavior specifies an unchanging probability of passing an item ($1/d_i$) regardless of examinee proficiency, rapid guesses add a constant, across all levels of θ, to the log-likelihood function and thus do not influence where the maximum of the function occurs regardless of the value of d_i. As a result, the item parameter estimates will be the same as those obtained from treating rapid guesses as not administered. In contrast, during item calibration for the standard 3PL model, item responses under rapid-guessing behavior were treated as valid.

During calibration for each of the IRT models, the *c*-parameter for each item was set equal to the reciprocal of the number of response options because the standard errors of estimated *c*-parameters would tend to be large with this sample size. Default priors were used on all item parameters. The posterior distribution of θ was estimated at each step with 15 quadrature points so that the scale was set by fixing the estimated posterior distribution to a mean of zero and standard deviation of one. During the calibrations, the estimates for three of the ILT items failed to converge. Consequently, these items were deleted from the analyses.

Results

Model Fit. To compare the fit of the two models, the likelihood of each response pattern of rights and wrongs was calculated based on each model and its respective item parameters. For the effort-moderated model, the single parameter for each item, d_i, was set equal to the item's number of response options. Following Levine and Drasgow (1988), the marginal likelihood of the response pattern was approximated using quadrature points. At each θ value, the likelihood of the response pattern was calculated and weighted by the proportion of examinees in that interval before averaging. The quadrature points covered the range of -3 to 3 at intervals of 0.1. After the marginal likelihoods were computed, the ratio of the likelihood based on the effort-moderated model to the likelihood based on the standard 3PL model was calculated for each person. This likelihood ratio approach has been used to compare alternative models by Drasgow, Levine, and McLaughlin (1987), Levine and Drasgow (1988), and Drasgow, Levine, and Zickar (1996).

Summary information for these ratios is shown in Table 1. For the examinees with RTE scores of 1.0, the only difference between the models was in the item parameters. For 83% of this group of examinees, the item parameters from the effort-moderated model fit their response patterns better than the item parameters from the standard 3PL model. In comparison, for the examinees who exhibited rapid-guessing behavior on one or more items, the models differed not only in their parameters but in the form of the model for those items on which the examinees rapid guessed. Across all examinees with *RTE* scores less than 1.0, the observed response patterns were more likely under the effort-moderated model 69% of the time. Thus, regardless whether or not an examinee exhibited rapid-guessing behavior on any items, most of the response patterns were more likely under the effort-moderated model.

Item Parameter Estimation. Figure 3 shows a scatter plot of the item difficulty parameters under the two models. For more difficult items (i.e., those with *b* parameters exceeding zero), the estimates were fairly similar for each model. For the remaining items there was a trend showing that, as the items got easier, the models tended to be in less agreement. The corresponding scatter plot for item discrimination is shown in Figure 4. In this graph, as discrimination increased, the models showed less agreement.

The direction and degree of item parameter disagreement between the models is summarized in Table 2. For each parameter, the values were ranked and divided into thirds. The mean parameter values are shown for all of the items as well for each third. For the discrimination parameter, the standard 3PL model yielded estimates that were, on average, 0.25 higher. More discriminating items, however, showed greater differences between the models, with a mean difference of 0.44. An interactive effect was also apparent for the difficulty parameter. Mean difficulty was substantially higher for the least difficult third of the items, and there were virtually no differences between the models for the most difficult third.

Test Information and Reliability. Figure 5 shows estimated test information functions based on the parameter estimates for each model. For the effort-moderated model, the information function is displayed only for those with *RTE* scores equal to 1.0; each examinee who exhibited one or more rapid-guessing behaviors would have a different information function because rapid guesses do not contribute to the test information. The information appears to be much greater for the standard 3PL model in the functional theta range between -2.0 to +2.0. This would be expected given the higher estimated discrimination parameters using the standard model.

Validity. Convergent validity information was obtained by generated proficiency estimates for examinees under each of the models and correlating them with external variables that would be expected to correlate positively with proficiency. These correlations, which are shown in Table 3, were consistently higher for the effort-moderated model.

Discussion

The results of Study 1 showed that, relative to the standard 3PL model, the effort-moderated model (a) fit the response patterns for far more examinees, (b) yielded substantially different item parameter estimates, (c) had lower estimated test information (which implies lower reliability), and (d) generated proficiency estimates with higher convergent validity. The finding that the effort-moderated model yielded proficiency estimates that were both less reliable and more valid seems contradictory. One explanation for these findings would be that the discrimination parameters from the standard model were spuriously high under the standard 3PL model due to the presence of the low-accuracy rapid guesses in the item response data. If this were the case, then test information would be spuriously high as well. Given that the effort-moderated model showed superior fit, the information function based on the effort-moderated model should be more accurate. This implies, therefore, that using the standard 3PL model in the presence of rapid-guessing behaviors can lead one to conclude that the scores are more reliable than they really are.

Study 2

In the sample of examinees investigated in Study 1, 29% of the sample had *RTE* scores below 1.0 and 5.8% of the total responses in the data set were rapid guesses. The differences observed between the effort-moderated and standard 3PL models would reasonably be expected to vary with the proportion of rapid guesses in a data set. To explore this further, Study 2 used simulated data with varying proportions of rapid guesses. Using simulated data also allowed the parameter estimates and estimated test information functions to be compared not just to each other but to known values. In Study 1, we reasoned that the estimates from the effort-moderated model were likely more accurate because the effort-moderated model fit the data better and the correlations with external criteria were higher using the proficiency estimates from the effort-moderated model. In Study 2, the estimated parameters and information were compared to the known values used to generate the data, so the accuracy of the estimates could be assessed directly. Therefore, the purposes of Study 2 were to:

1. Determine whether the item parameter estimates and test information functions were more accurate under the effort-moderated model.
2. Explore how changes in the proportion of responses that were rapid guesses affected the accuracy of the effort-moderated and standard 3PL models.

Method

Data for Study 2 were based on the item parameters and scores from Study 1. To vary the proportion of examinees exhibiting rapid-guessing behavior, the percent of simulees with *RTE* scores of 1.0 was set to either 50%, 70%, or 90%. The remaining simulees in each of the three conditions had *RTE* scores that were distributed in a manner proportional to those found with the data in Study 1. In addition to keeping the

distribution of *RTE* scores of rapid guessers proportional to the distribution of *RTE* scores of rapid guessers in the real data set (while varying the overall percent of rapid guessers), it was desirable to maintain the complex relationship between θ and *RTE* score. This was not simply a linear relationship; there was a wide range of θ for those with *RTE* scores of 1.0, but for other *RTE* scores there was a negative relationship with θ. In order to replicate this relationship, a sampling with replacement or bootstrap method was used to sample θ's and *RTE* scores from the Study 1 data. For each replication, first 50%, 70%, or 90% of the sample was drawn, with replacement, from the group with *RTE* scores of 1.0. Then the remainder of the sample was drawn, again with replacement, from the group with *RTE* scores of less than 1.0. For each of the 100 replications, 500 simulees were sampled, which was approximately equal to the sample size in Study 1.

After the θ's and *RTE* scores were sampled, item responses were generated. The item parameter estimates from the effort-moderated model in Study 1 were used as the true item parameters to generate the data. Based on a simulee's *RTE* score, the number of items that would be rapid guesses was determined; these items were randomly selected from the 57 items. Then the probability of correct response for each item was calculated using Equation 4; if a number drawn randomly from a uniform[0,1] distribution was lower than this probability, the response was coded correct.

Item parameters were again calibrated using BILOG-MG 3, using the same options used in Study 1. Each data set was calibrated once using the standard 3PL model and again using the effort-moderated model. For the standard 3PL model, all responses were considered valid; for the effort-moderated model the rapid guesses were coded as not administered. Note that in Study 2 it was known whether the item response was a rapid guess or a valid response; this did not need to be estimated from the response time as in Study 1. A few items had negative item-total correlations in some replications and were automatically omitted from the calibration (6 of the 5700 items using the standard 3PL model and 7 items using the effort-moderated model).

Because θ was correlated with response time effort, the distribution of the true (generating) θ's varied with the condition. On average, the true θ's had a mean of 0.21 and standard deviation of 0.97 in the 90% condition, -0.05 and 1.16 in the 70% condition, and -0.32 and 1.25 in the 50% condition. In BILOG-MG, the *empirical* option was used to scale the parameters such that the posterior mean of θ was 0 and the standard deviation was 1. Thus, the estimated parameters needed to be equated before comparison with the true parameters. For each replication, each θ estimate and item difficulty estimate was multiplied by the standard deviation of the true θ's sampled for that replication, and the mean of the true θ's was added to each θ estimate and each item difficulty estimate. Then, each item discrimination estimate was divided by the standard deviation of the true θ's. In other words, the scale was adjusted such that the estimated posterior distribution had a mean and standard deviation equal to that of the sample of true θ's used in that replication. Alternatively, the item difficulties could have been equated, but in this context it made more sense to equate the θ's so that changes in item difficulty estimates could be studied.

Results

In the 90% condition, 2.3% of the responses were rapid guesses. In the 70% condition, 6.7% of the responses were rapid guesses; in the 50% condition, 11.3% of the responses were rapid guesses.

Item Parameter Recovery. After equating the item parameters as described in the *Method* section, the average bias and root mean square error (RMSE) was calculated for each item, and then averaged across items. These averages for bias and RMSE are displayed in Tables 4 and 5, respectively. Both unweighted means and weighted means, using the inverse of the estimated squared standard error as the weight, are included. The weighted means are less impacted by extreme parameter estimates because extreme estimates have larger standard errors. This is particularly evident for the *b* parameters, because the very easy items tended to have the most bias. Overall, both the *a*'s and the *b*'s were positively biased using the standard 3PL model, meaning that the items appeared to be more difficult and more discriminating than they really were. The parameter estimates from the effort-moderated model were slightly negatively biased, but the absolute value of the bias was smaller than with the standard model. For both models, the absolute value of the bias increased as the proportion of rapid guessers increased. The RMSE was also smaller for the effort-moderated model. This would be expected, given the smaller bias, unless the parameter estimates had been much more variable using the effort-moderated model due to the reduced effective sample size. The RMSE, however, followed the same pattern as the bias.

Because the bias results were highly similar for the 50% and 70% simulation conditions and the RMSE results were highly consistent with the corresponding bias results, only scatter plots of the bias results for the 70% and 90% conditions are presented. Figures 6 and 7 show the bias in estimating the discrimination parameters for the 70% and 90% conditions, respectively. In each graph, the differences between the discrimination estimates based on the two models remained fairly constant across the range of discriminations, though the differences were somewhat larger for the more highly discriminating items (as was found in Study 1).

The bias in estimation of the item difficulty parameters also followed the pattern found in Study 1.The respective scatter plots for the 70% and 90% conditions are shown in Figures 8 and 9. In each graph, the easiest items were positively biased for the standard 3PL model, and to a lesser extent negatively biased for the effort-moderated model (except for the *extremely* easy items with difficulties near -4, which were likely positively biased due to the use of a prior distribution in the calibration). In contrast, the most difficult items had a small negative bias for the standard model and a small positive bias for the effort-moderated model.

Test Information. The estimated information functions for the 70% and 90% conditions, based on the mean estimated parameters for each item[2], along with the true test information are graphed in Figures 10 and 11, respectively. The information functions for the effort-moderated models were calculated only for examinees with *RTE*

scores of 1.0; rapid guesses would contribute nothing to a given information function so each examinee exhibiting rapid-guessing behavior on at least one item would have a different information function depending on which items were rapid guesses. The results for the 70% and 90% simulation conditions were consistent; for most examinees (i.e., those with proficiency between -2.0 and +2.0), test information was clearly overestimated by the standard 3PL model, while being slightly underestimated by the effort-moderated model. The information differences between the models increased as the proportion of examinees showing *RTE* scores equal to 1.0 decreased, but were clearly still apparent in the 90% condition (in which only 2.3% of the responses were rapid guesses).

Discussion

The simulation study confirmed our hypothesis that the effort-moderated model item parameters were more accurate. They were less biased and had smaller RMSEs. Using the standard 3PL models, the items appeared more difficult and more discriminating than they actually were. Across simulation conditions, due largely to their more accurate parameter estimation, the effort-moderated models also better reproduced the true test information functions. Overall, the results of Study 2 support the conclusion that the effort-moderated model more effectively models examinee behavior when the data are influenced by even a modest amount of rapid-guessing behavior.

The results of this study are consistent with those found in previous research. Oshima (1994) studied item parameter estimation under speeded conditions, finding that under speeded conditions (which would elicit rapid-guessing behavior), the standard 3PL model showed inflated test information, and overestimated the *a* and *b* parameters. Similarly, Bovaird (2002) found that the removal of rapid-guessing examinees from test data resulted in a significant decrease in item discrimination, though he observed a significant increase in item difficulty.

A particularly interesting Study 2 finding was that the presence of rapid-guessing behavior negatively affected the performance of the standard 3PL model, even in the 90% condition. This suggests that even when the preponderance of rapid-guessing behavior was fairly small (i.e., half of that found in Study 1), the effort-moderated model exhibited clearly more desirable psychometric characteristics than did the standard 3PL model.

General Discussion

A general goal of the measurement professional is to obtain valid test scores. Whenever test performance has little or no consequences for examinees, however, it is likely that some examinees will direct little effort toward their tests. This represents a direct threat to test score validity, and measurement professionals have a responsibility for taking corrective action. This paper introduces a new method for modeling the presence of rapid-guessing behaviors—which have been shown to be indicators of low examinee effort (Wise & Kong, in press)—that measurement professionals might effectively use to manage this problem.

The results of the two studies reported here suggest that the effort-moderated model would be more appropriate than standard IRT models when even a modest amount of rapid-guessing behaviors is present (e.g., 2% or more). If one can assume that the computerization of a paper-and-pencil test does not elicit additional rapid-guessing behavior, this represents an important additional advantage of computer-based tests. Directors of low-stakes testing programs should therefore consider the utility of being able to measure item response time, which could be used either to merely monitor the effort examinees give toward their tests (i.e., using *RTE* scores), remove the data from examinees who exhibited low effort (i.e., motivation filtering), or to incorporate response time into proficiency estimation using an effort-moderated IRT model.

The effort-moderated model investigated in Studies 1 and 2 is compatible with commonly-used IRT software. After the rapid guesses are identified, the model can be implemented in BILOG or any software that allows items to be coded as not administered. Because rapid guesses add a constant to the likelihood function, they do not influence where the likelihood function for an item or examinee peaks. This can also be viewed mathematically; the maximum occurs where the first derivative is zero, and a constant does not change the first derivative. Thus, when estimating *a*, *b*, *c*, or θ, there is no need to consider the value of *d*. The value of *d*, though, may be desirable for modeling the probability of the response given by a particular examinee. This was necessary for the model fit procedure used in Study 1, for example. If a large number of examinees exhibited rapid-guessing behaviors to an item, *d* could be estimated empirically as the proportion of rapid guesses that were correct. A more simple method was used here; *d* was set equal to 1/(the number of options) because an examinee who makes a rapid guess would not be expected to have time to eliminate obviously wrong distractors and would guess randomly from all options.

It should be noted that the effort-moderated model investigated in this paper was applied in all instances of rapid-guessing behavior, regardless of how often this had occurred for a given examinee. In practice, however, one may want to require a minimum percentage of solution behaviors be exhibited for an examinee to obtain a score. One might, for example, set the minimum at 75%, which implies that any examinee with an *RTE* score less than 0.75 would not receive a valid score, as there would be insufficient information for that examinee's score to be credible.

Wise (in press) stated that, with low-stakes tests, measurement professionals have a responsibility to either develop methods for identifying and managing data from examinees who do not give good effort or to adopt testing practices that promote examinee effort. Using computer-based testing and employing the effort-moderated model exemplifies the first strategy. The second strategy can be realized through the use of relative short tests, with limited amounts of reading, and that are intrinsically motivating to examinees. A combination of these two strategies, however, might be particularly effective for measurement professional to use.

References

Bovaird, J. A. (2002). New applications in testing: Using response time to increase the construct validity of a latent trait estimate. *Dissertation Abstracts International*. (UMI No. 3082643).

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59-79.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9*, 47-64.

Eccles, J. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 75-146). San Francisco: Freeman.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*, 133-146.

Lau, S. & Roeser, R. W. (2002). Cognitive abilities and motivational processes in high school students' situational engagement and achievement in science. *Educational Assessment, 8*, 139-162.

Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53*, 161-176.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*, 200-219.

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications* (2nd ed.). Upper Saddle, NJ: Merrill Prentice-Hall.

Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Schnipke, D. L. (1996, April). *How contaminated by guessing are item-parameter estimates and what can be done about it?* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY. (ERIC Document Reproduction Service No. ED400276)

Schnipke, D. L. (1999). *The influence of speededness on item-parameter estimation* (Computerized Testing Report No. 96-07). Princeton, NJ: Law School Admission Council. (ERIC Document Reproduction Service No. ED467809)

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.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In Mills, C. N., Potenza, M.T., Fremer, J. J., & Ward, W. C. (Eds.). *Computer-based testing: Building the foundation for future assessments.* Mahwah, NJ: Lawrence Erlbaum Associates.

Snow, R. E. (1989). Cognitive-conative aptitude interactions in learning. In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation, and methodology* (pp. 435-474). Hillsdale, NJ:Lawrence Erlbaum Associates, Inc.

Snow, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational Psychologist, 27*, 5-32.

Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wagner (Eds.) *Mind in context: Interactionist perspectives on human intelligence* (pp. 3-37). Cambridge: Cambridge University Press.

Sundre, D. L., & Wise, S. L. (2003, April). *'Motivation filtering': An exploration of the impact of low examinee motivation on the psychometric quality of tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Tatsuoka, K. K. (1996). Use of generalized person-fit indices, zetas for statistical pattern classification. *Applied Measurement in Education, 9*, 65-76.

Wigfield, A., & Eccles, J. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68-81.

Wise, S. L. (2004). An investigation of the differential effort received by items on low-stakes, computer-based tests. Manuscript under review.

Wise, S. L., & DeMars, C. E. (in press). Examinee motivation in low-stakes assessment: Problems and potential solutions. *Educational Assessment*.

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

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3.0* [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

Acknowledgement

The authors wish to thank Xiaojing Kong both for her assistance with the data analyses and for her insightful comments and suggestions regarding the manuscript.

Footnotes

[1] For example, the visually-chosen threshold for the item depicted in Figure 1 would be about four seconds.

[2] Alternatively, the information function could have been estimated for each replication and then averaged. This would potentially lead to an average function that was flatter and more disperse than the typical within-replication information function.

Table 1

Ratio of Likelihood of Response Patterns Under Effort-Moderated Model to Likelihood Under the Standard 3PL IRT Model

Response Time Effort Value	*N*	Median Likelihood Ratio	Percent of Ratios Exceeding 1.0
1.00	386	1.25	83
.900 - .999	74	1.47	68
.750 - .899	24	7.38	62
.500 - .749	14	33.16	100
< .500	26	7.25	62

Table 2

Mean Discrimination and Difficulty Parameter Estimates Under the Standard 3PL and Effort-Moderated IRT Models

Item Group	Number of Items	Mean Parameter Value		Mean Difference
		Standard 3PL	Effort-Moderated	
		Discrimination Parameter		
Least Discriminating Third	19	0.58	0.47	0.11
Middle Third	19	0.82	0.62	0.20
Most Discriminating Third	19	1.25	0.81	0.44
All Items	57	0.88	0.63	0.25
		Difficulty Parameter		
Least Difficult Third	19	-1.75	-2.54	0.79
Middle Third	19	-0.54	-0.99	0.45
Most Difficult Third	19	0.77	0.84	-0.07
All Items	57	-0.51	-0.90	0.39

Table 3

Convergent Validity Correlations Between ILT Estimated Proficiency and Several External Variables for Each IRT Model

	IRT Model	
External Variable	Standard 3PL	Effort-Moderated
SAT-Verbal	.33	.40
SAT-Quantitative	.13	.19
Grade Point Average	.24	.27

Note. All correlations were statistically significant at the .01 level. $N = 488$.

Table 4

Bias in Item Parameters for the Three Simulated Conditions

Parameter	Percent of Simulees with RTE = 1.0	Unweighted Mean		Weighted Mean	
		Standard 3PL	Effort-Moderated	Standard 3PL	Effort-Moderated
a					
	50	0.17	-0.05	0.15	-0.06
	70	0.17	-0.04	0.14	-0.05
	90	0.17	-0.01	0.10	-0.03
b					
	50	0.44	-0.10	0.30	-0.08
	70	0.40	-0.08	0.23	-0.04
	90	0.31	-0.01	0.15	0.01

Note. The percentages of rapid guesses in the 50%, 70% and 90% simulation conditions were 11.3, 6.7, and 2.3, respectively.

Table 5

RMSE in Item Parameters for the Three Simulated Conditions

Parameter	Percent of Simulees with RTE = 1.0	Unweighted Mean		Weighted Mean	
		Standard 3PL	Effort-Moderated	Standard 3PL	Effort-Moderated
a					
	50	0.24	0.14	0.21	0.13
	70	0.24	0.14	0.20	0.12
	90	0.22	0.14	0.17	0.13
b					
	50	0.71	0.53	0.52	0.24
	70	0.67	0.41	0.45	0.23
	90	0.58	0.44	0.34	0.24

Note. The percentages of rapid guess in the 50%, 70% and 90% simulation conditions were 11, 7, and 2, respectively.

Figure Captions

Figure 1. Distribution of examinee response times for an information literacy test item.

Figure 2. An example of effort-moderated item response functions.

Figure 3. Scatter plot of the item difficulty parameters for the standard 3PL and effort-moderated IRT models.

Figure 4. Scatter plot of the item discrimination parameters for the standard 3PL and effort-moderated IRT models.

Figure 5. ILT test information functions for the standard 3PL and effort-moderated IRT models.

Figure 6. Bias in the estimation of the *a* parameters under the two models for the simulated 70% condition.

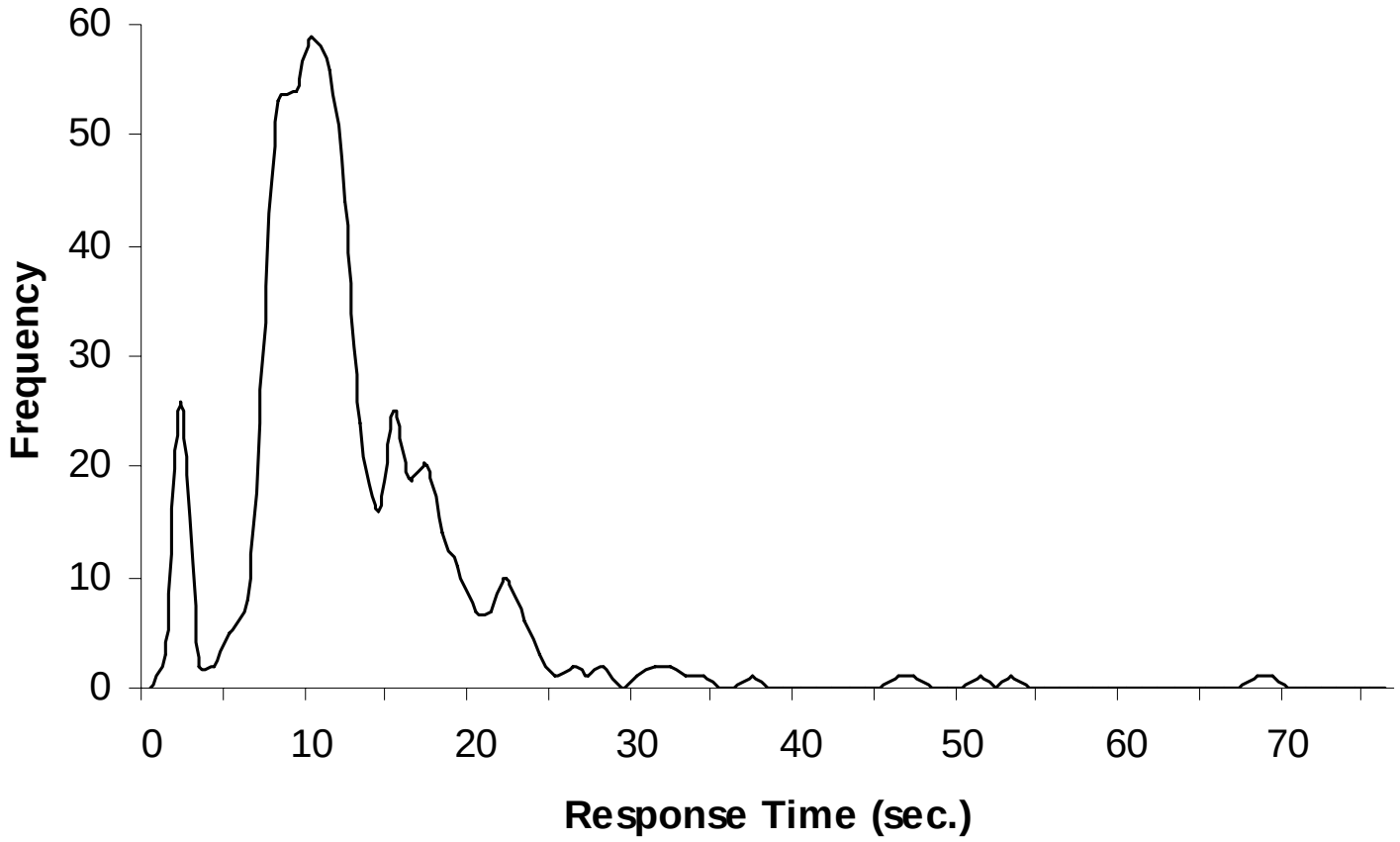
Figure 7. Bias in the estimation of the *a* parameters under the two models for the simulated 90% condition.

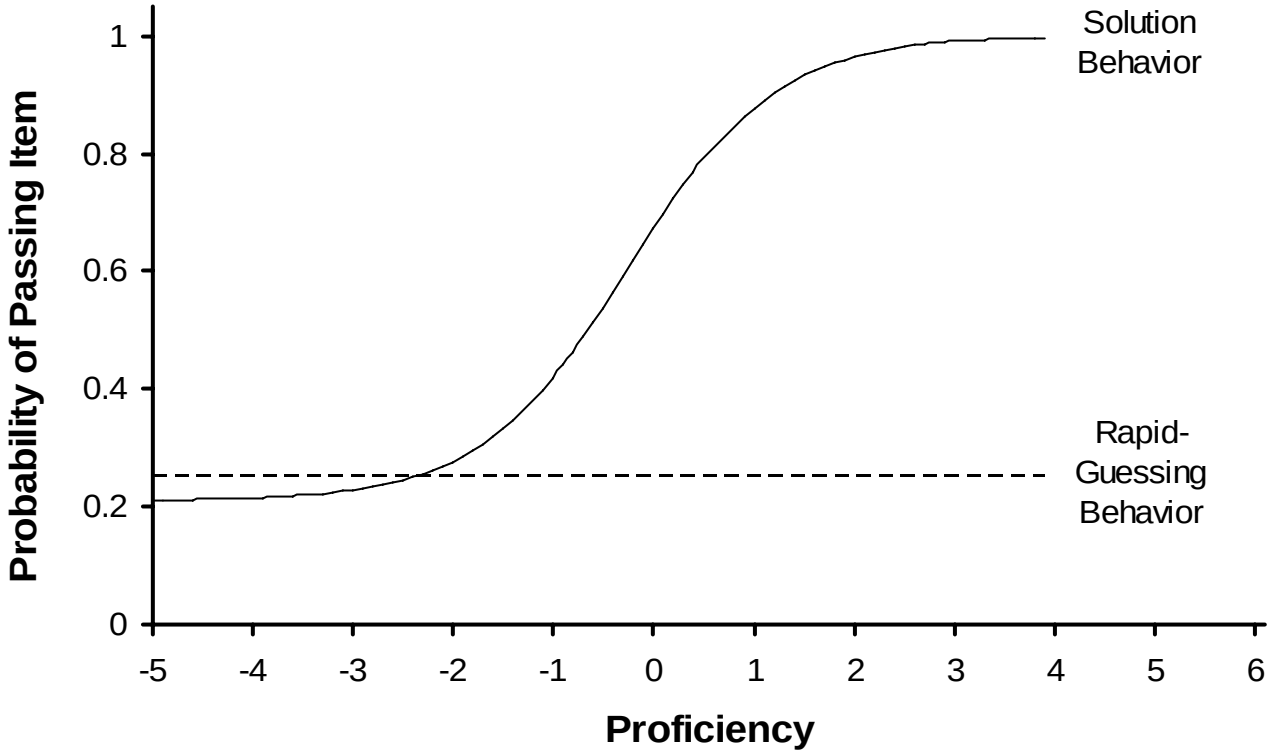
Figure 8. Bias in the estimation of the *b* parameters under the two models for the simulated 70% condition.

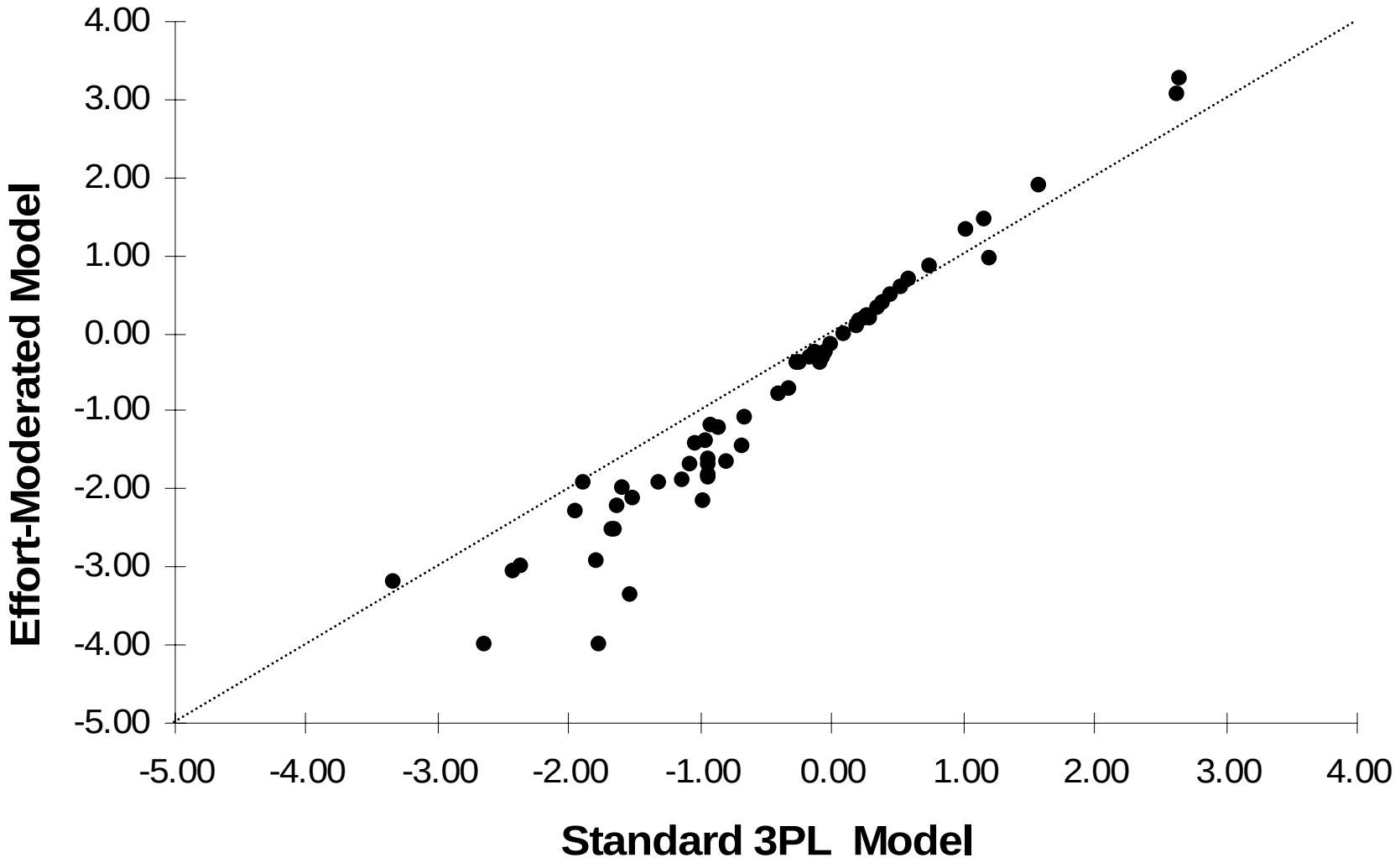
Figure 9. Bias in the estimation of the *b* parameters under the two models for the simulated 90% condition.

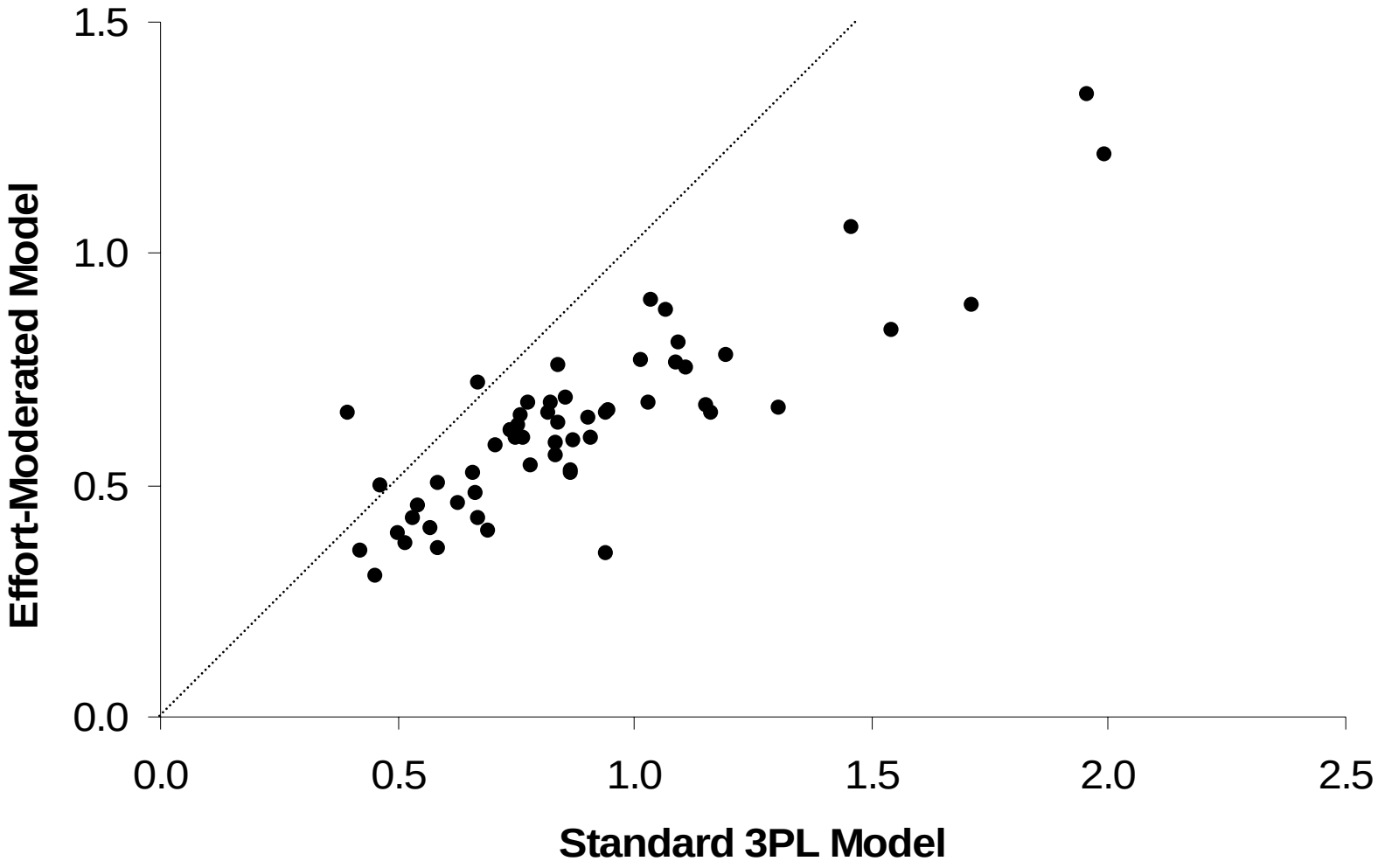
Figure 10. Test information functions for the standard 3PL and effort-moderated models (simulated 70% condition) along with true test information.

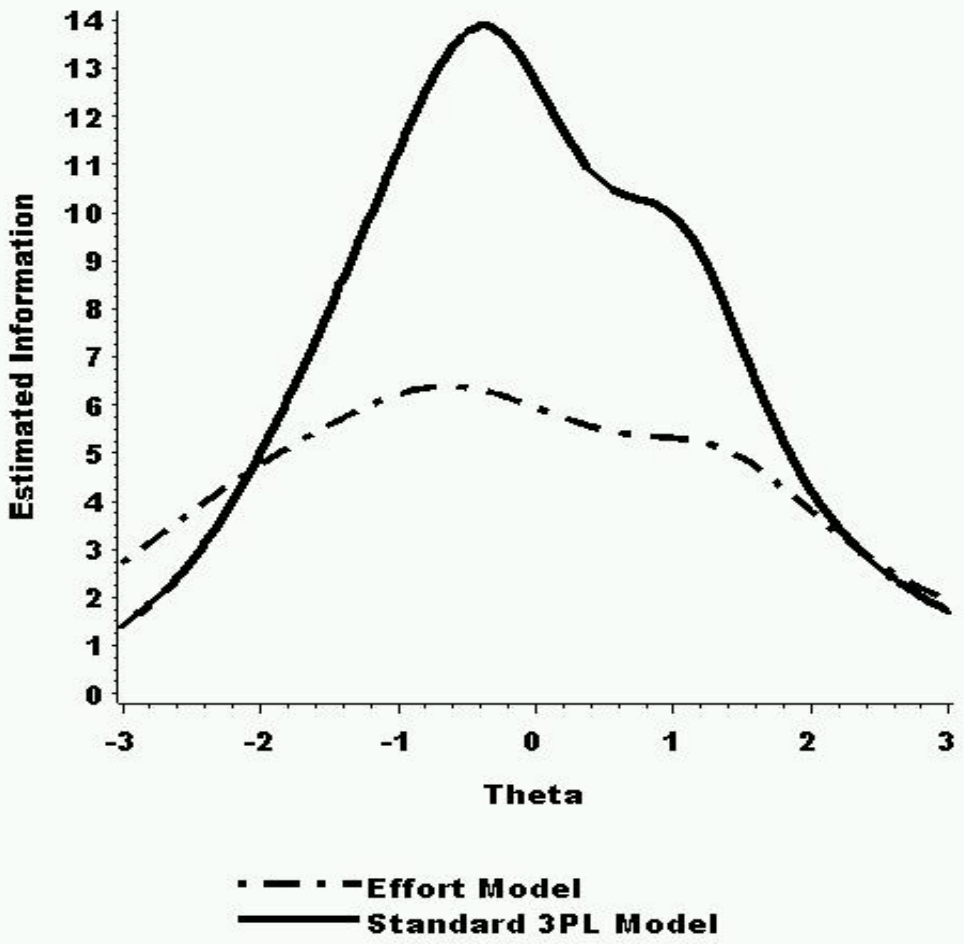
Figure 11. Test information functions for the standard 3PL and effort-moderated models (simulated 90% condition) along with true test information.

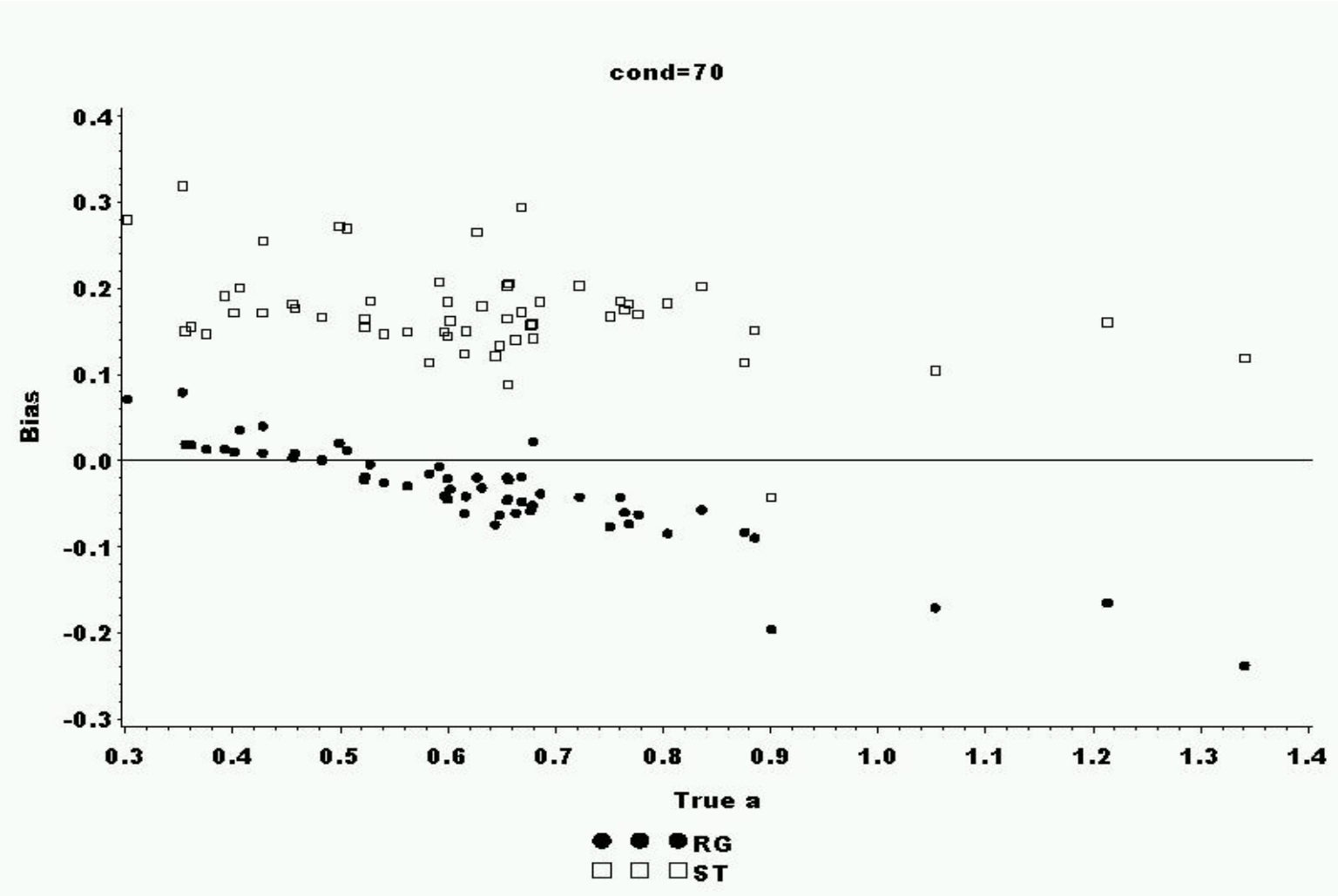

Frequency
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
Response Time (sec.)

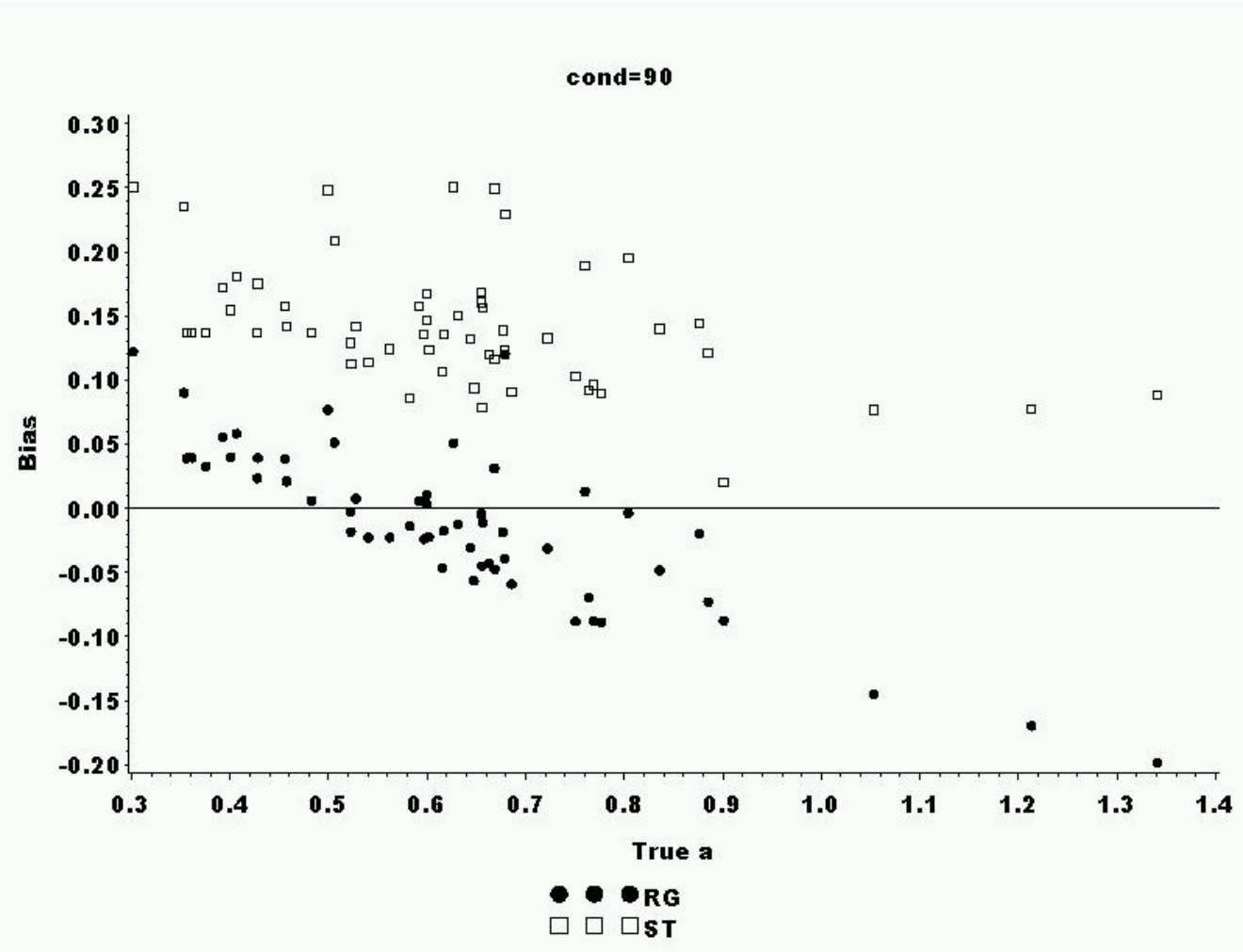

Probability of Passing Item
1
0.8
0.6
0.4
0.2
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
Proficiency
Solution Behavior
Rapid-Guessing Behavior

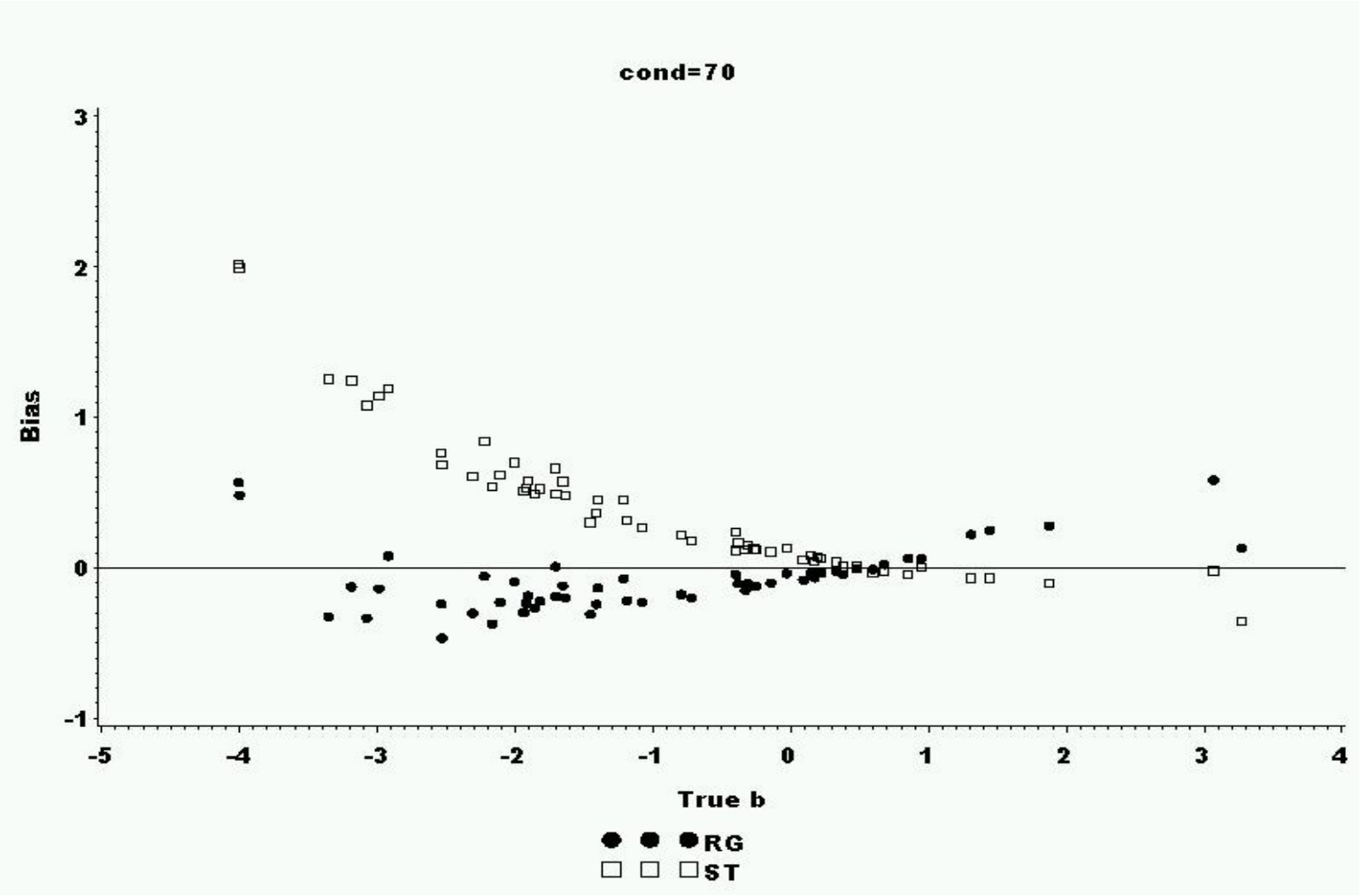

Effort-Moderated Model
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
-4.00
-5.00
-5.00
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
3.00
4.00
Standard 3PL Model

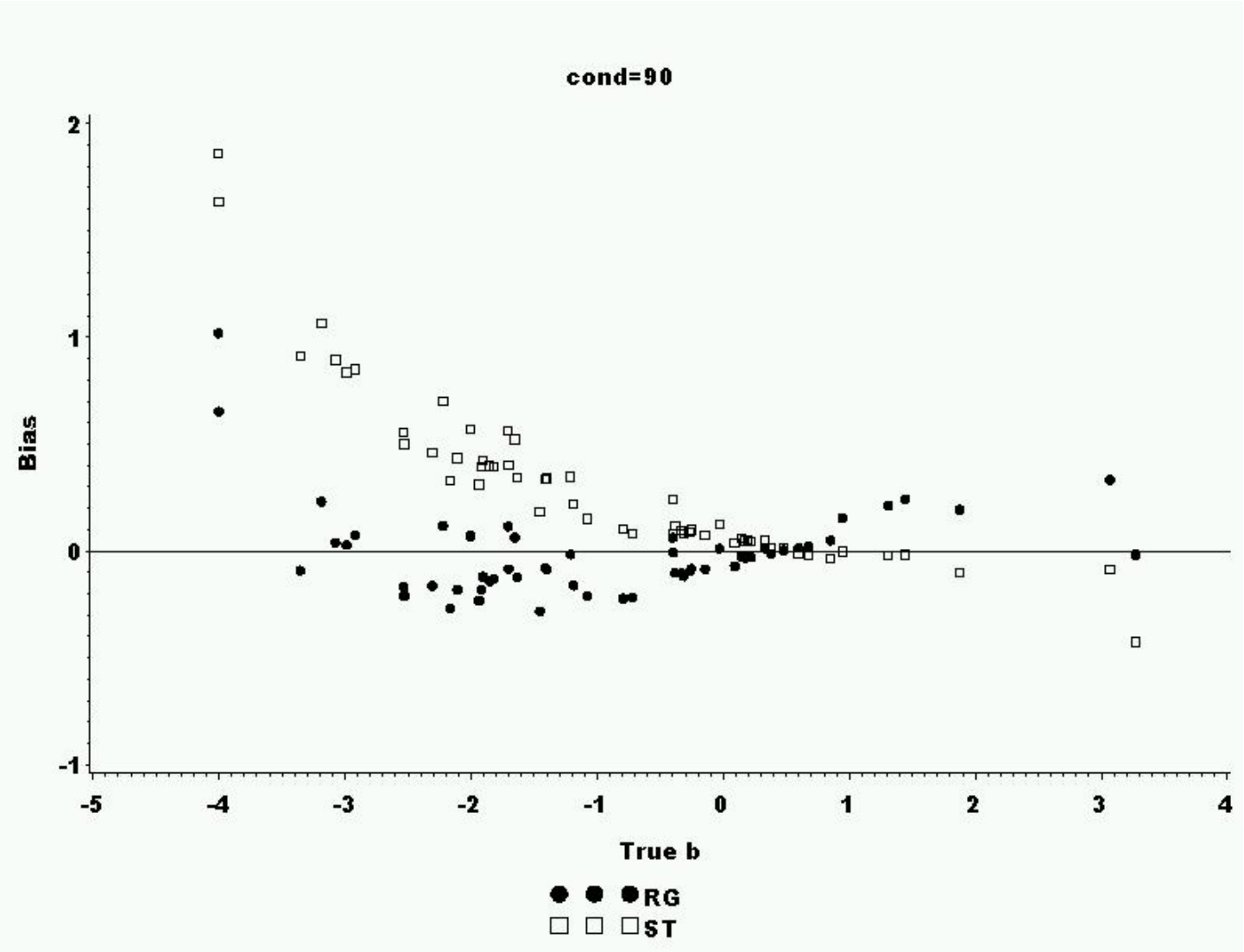

Effort-Moderated Model
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
Standard 3PL Model

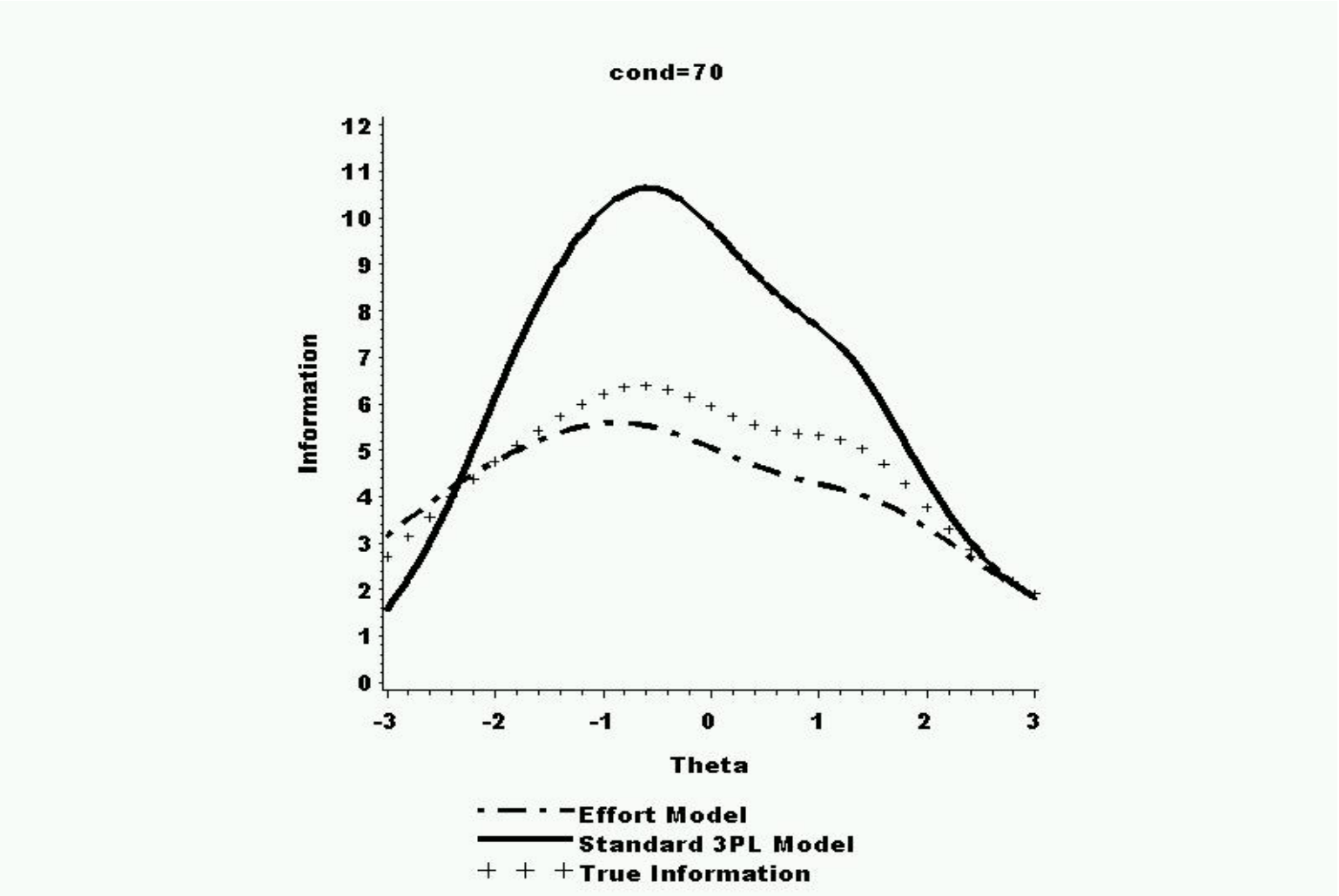

Estimated Information
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
-3
-2
-1
0
1
2
3
Theta
Effort Model
Standard 3PL Model

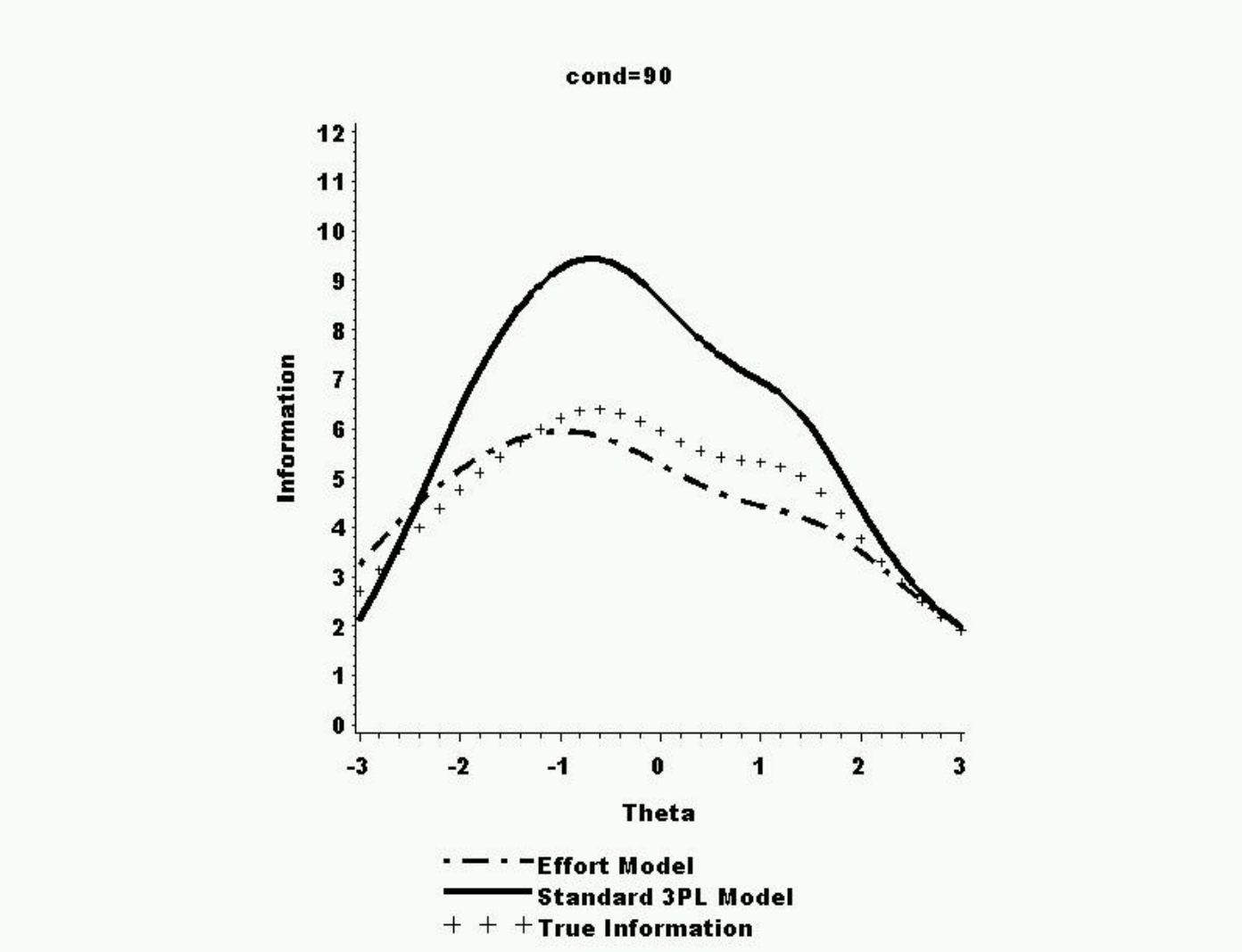

cond=70
Bias
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
True a
RG
ST


cond=90
Bias
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
True a
RG
ST


cond=70
Bias
True b
RG
ST
3
2
1
0
-1
-5
-4
-3
-2
-1
0
1
2
3
4


cond=90
Bias
2
1
0
-1
-5
-4
-3
-2
-1
0
1
2
3
4
True b
RG
ST


cond=70
Information
12
11
10
9
8
7
6
5
4
3
2
1
0
-3
-2
-1
0
1
2
3
Theta
Effort Model
Standard 3PL Model
True Information


cond=90
Information
12
11
10
9
8
7
6
5
4
3
2
1
0
-3
-2
-1
0
1
2
3
Theta
Effort Model
Standard 3PL Model
True Information