Running head: STUDYING THE EFFECT OF RAPID-GUESSING

# Studying the Effect of Rapid-Guessing on a Low-Stakes Test: An Application of the Effort-Moderated IRT Model

J. Carl Setzer

Center for Assessment and Research Studies

James Madison University

Jill R. Allspach

Educational Testing Service

October 18, 2007

# Studying the Effect of Rapid-Guessing on a Low-Stakes Test: An Application of the Effort-Moderated IRT Model

One of the assumptions often made in testing and assessment situations is that examinees engage the test items with full effort. This assumption, whether made for convenience or with naiveté, may result in misleading estimates of item and/or person characteristics. This is particularly the case in low-stakes testing situations, where the test scores are of no personal consequence to the examinee. In fact, empirical research has found support for claims that test score interpretations are less valid in low-stakes compared to high stakes situations (Wise & DeMars, 2003). Given that no consequences are attached to the final test score at the examinee level, it should surprise no one that test-takers do indeed vary in effort levels.

Several methods have been developed to either measure or model the effort levels in examinees. For example, a popular method for measuring effort is to simply ask examinees about their individual effort on the test. Often, simple questions such as “how hard did you try on the test you just took?” or “how well did you want to do on this test?” are used (Wise, Kingsbury, Thomason, and Kong, 2004). Other, more formally developed scales are also used, such as one developed by Wolf and Smith (1995), later revised and coined the Student Opinion Survey (SOS) by Sundre (1999). An obvious issue surrounding the use of such measures of effort is the subjective nature of the measurement process. An examinee who just minutes before completed a test with little effort has at least some propensity to put little effort into answering follow-up questions. Moreover, some examinees no doubt tend to “sabotage” the validity of their test scores by claiming they put less effort into their responses than they actually did. That is, they may be trying to demonstrate that they *could* have done better on the test if their effort level was higher.

Another more objective method of measuring effort, used mainly in conjunction with computer-based testing (CBT), involves the study of response times. Response times can be collected at both the item level and across all items (i.e., total test time). Presumably, low-motivated examinees spend less time contemplating their responses to items. Consequently, considerably lower response times at both the item and test levels could act as a red flag.

Wise (2006) showed that effort levels may wax and wane throughout a non-speeded test. To assess effort at both the item- and test-level, Wise and Kong (2005) devised a general measure called *Response Time Effort* (RTE) that is based on the notions of *solution behavior* and *rapid-guessing behavior* (Schnipke, 1995, 1996, 1999; Schnipke & Scrams, 1997, 2002). Solution behavior (SB) refers to the situation in which an examinee puts at least a minimum amount of effort into their item response whereas rapid-guessing behavior (RGB) indicates that the effort level is insufficient. Classifying an examinee’s response as being the result of either SB or RGB is determined at the item level. RTE is simply the sum of the SB across all items for examinee *i* as follows:

$$RTE_j = \frac{\sum SB_{ij}}{k} \qquad (1)$$

In equation 1, *k* is equal to the number of items on the test (*j*=1, 2,…,*k*). RTE, then, represents the proportion of items in which examinee *i* put forth a minimal level of effort.

Prior to distinguishing between SB and RGB, a threshold value must be established for each test item. Various methods for setting these thresholds have been examined. These methods include using a common threshold for all items, using characteristics of the items (e.g., number of words in the stem), examining the response time distributions, and using multi-state mixture modeling. However, Kong, Wise, and Bhola (2007) examined each of these methods empirically and found few differences. These authors concluded that a simple examination of the response time distribution should provide a practical and effective means for determining the threshold values.

Although analyzing the response times can be altogether interesting, we are also seeing more and more studies that include these variables in more advanced modeling techniques. Recently, Wise and DeMars (2006) introduced the effort-moderated item response model (EMIRM) in an attempt to compensate for RGB at the item level. Essentially, the EMIRM treats responses obtained via RGB as “not administered”. The EMIRM has been shown via simulation to result in more accurate test information functions when compared to the standard three-parameter IRT model. The model incorporates a variable for SB as follows:

$$P_i(\theta) = (SB_{ij})\left(c_i + (1-c_i)\left(\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}\right)\right) + (1-SB_{ij})(g_i) \qquad (2)$$

In equation 2, $SB_{ij}$ takes a value of 1 if the response time is greater than the set threshold and 0 otherwise. When $SB_{ij} = 1$ then the standard 3PLM is applied. Otherwise, when $SB_{ij} = 0$ the probability of getting the item correct is simply a function of $g_i$. In the EMIRM, $g_i$ is the reciprocal of the number of response options for item *i*. As described by Wise and DeMars, when RGB occurs the probability of getting the item correct should be at approximately chance level, regardless of ability. Using $g_i$ in place of the standard 3PL model when RGB occurs simply adds a constant to the log-likelihood function for each level of ability.

The purpose of this paper is to employ the techniques introduced by Wise and DeMars (2006) and to demonstrate their use on a large-scale, nationally recognized collegiate-level test. No theoretical advances are presented in this paper. Rather, the intent is to bring greater attention to the potential effects caused by low-effort and to the analytic techniques that have recently been developed to compensate for such effects.

## Method

*Examinees.* The sample used in this study consisted of 10,104 examinees. Of this sample, approximately 54 percent were males; 80 percent were White/Caucasian and approximately 7 percent were Black or African American. All examinees were in their senior year at their respective college or university.

*Assessment test.* The current study used data from one of the Major Field Tests developed by ETS. The test is designed to measure the examinee's knowledge of the subject and requires application of theory, concepts, facts, and analytical methods. The test is comprised of 120 multiple choice, dichotomously-scored items which are administered in two sections of 60 items each. All items require a response and therefore no missing data were included in the data file. The test is administered either as a paper-and-pencil version or a computer-based version. Only data from the computer-based version were used in this study. The computer-based version of the test records the response time (in whole seconds) for each examinee-by-item interaction.

Currently, ETS uses a Classical Test Theory framework to develop and score its Major Field Tests. In order to apply an IRT model to the data, it was first necessary to determine whether the data are essentially unidimensional. To make this assessment, we used DIMTEST 2.0, which tests the null hypothesis that the test is essentially unidimensional. The test was split into two sets of items (i.e., items 1-60 in the first set and items 61-120 in the second set). The dataset was then split into two subsets, each representing approximately 50 percent of the total sample. Each set was assigned to either the first or second set of test items. Finally, DIMTEST was used to assess the dimensionality of each test half. The results suggested that the first 60 items were essentially unidimensional (*Stout's t* = 0.3114; *p* = 0.3778) while the second set of items were not (*Stout's t* = 3.8401; *p* < 0.001). The analysis proceeded using the first 60 items on the test and the full sample of examinees.[1]

Thresholds for assessing SB were determined by examining the response time distribution for each item. The threshold was placed at the point on the scale where there appeared to be an intersection between a rapid-guessing distribution and a solution-behavior distribution. For example, Figure 1 displays the frequency of response times for an item from the test. The distribution appears to be bimodal with a minor distribution that occurs between the values of approximately 0 and 6 seconds and another larger distribution that begins around 6 seconds and continues to the maximum response time. In this case, then, it could be argued that the threshold should be set at 6 seconds. Examinees who spent 6 seconds or less on this particular item would be classified as rapid-guessers. The thresholds for all remaining items were obtained in a similar manner.

[1] The null hypothesis of essential unidimensionality was also rejected for the entire set of test items.

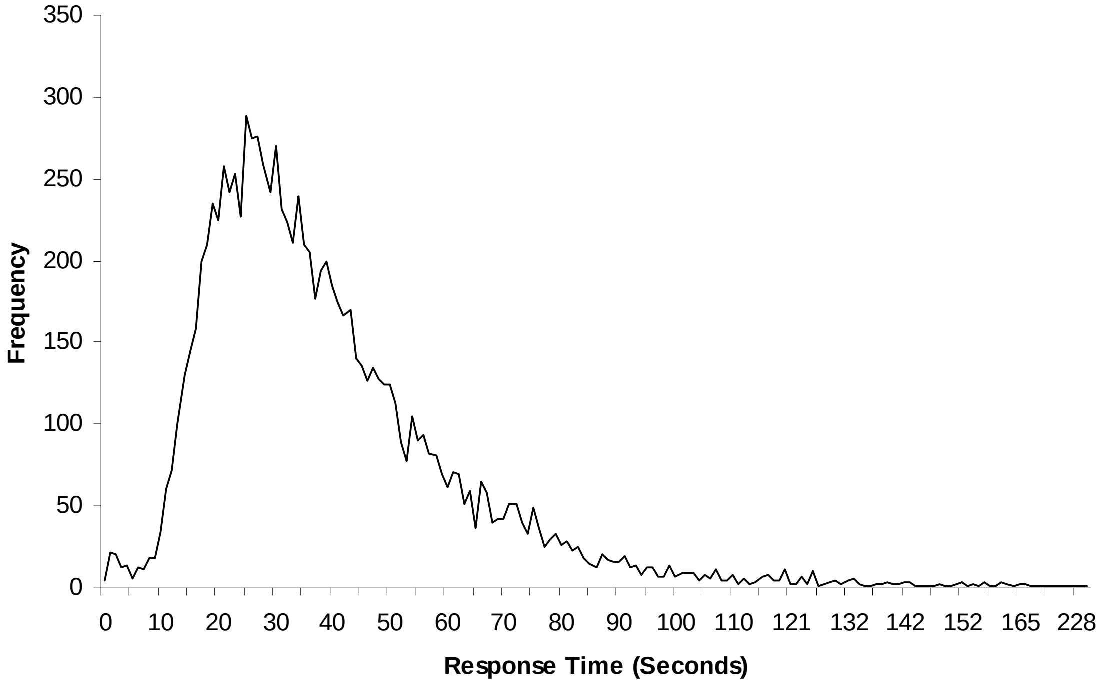

FIGURE 1. *Distribution of examinee response times for an item from the Major Field Test Data.*

*IRT calibration.* Both the standard 3PL and EMIRM were calibrated using the MFT data described above. For the EMIRM, responses obtained via rapid-guessing were recoded as "not administered" for calibration. Both models were calibrated using BILOG-MG 3 (Zimowski, Muraki, Mislevy, & Bock, 2003). Due to the large sample size and test length, standard marginal maximum likelihood estimation was used (i.e., Bayesian priors were not applied to any of the item parameters during either calibration).

## Results

*Model fit.* The standard 3PL model and the EMIRM are non-nested models. However, given both models' -2loglikelihood values (i.e., model deviances) were positive and that both IRT models contained the same number of estimated item parameters (in addition to the fact that the sample sizes were similar) it was not necessary to calculate either the AIC or BIC values. Instead, we simply compared the -2LL values obtained after the final iteration. The -2LL for the EMIRM was 729,904 while the -2LL for the standard 3PL was 736,190. In this case the model deviance is smaller for the EMIRM, suggesting better fit relative to the standard 3PL model.[2]

To further compare the relative fit of the models, we first calculated the likelihood of the response patterns given each model. Next, the ratios of these likelihoods were calculated such that a value greater than 1.0 indicated better fit with the EMIRM. Table 1 summarizes these ratios across levels of RTE. For example, the fit of the two models was relatively similar when RTE was equal to 1.0 (this represented the vast majority of the response patterns). However, as RTE decreases, it appears that the relative fit of the models tends to favor the EMIRM. In fact, 95 percent of the ratios calculated favored the EMIRM when RTE was less than 0.5.

Table 1

*Ratio of Response Pattern Likelihoods obtained via the Effort-Moderated and Standard 3PL IRT Models*

| Response Time Effort Value | *N* | Median Likelihood Ratio | Percent of Ratios Exceeding 1.0 |
|---|---|---|---|
| 1.00 | 9,162 | 1.01 | 61 |
| .900-.999 | 733 | 1.04 | 56 |
| .750-.899 | 93 | 2.86 | 73 |
| .500-.749 | 54 | 27.43 | 80 |
| <.500 | 62 | 1731.60 | 95 |

[2] For reference, the AIC values were 730,144 and 736,430 and the BIC values were 731,010 and 737,296 for the EMIRM and standard 3PL IRT models, respectively.

*Item parameter estimates.* Figures 2 through 4 display scatterplots of the estimated difficulty, discrimination, and pseudo-guessing parameters, respectively. An inspection of these plots suggests that the difficulty and discrimination parameter estimates are quite similar across the models. In fact, the correlations for both parameters are greater than .99. Less congruence between the estimates can be seen with the pseudo-guessing parameter estimates. Although the correlation here is .92, there does not seem to be much of a discernable pattern of bias (i.e., the models tended to disagree across the entire range of *c* values). Note that there were four items estimated to have a lower-asymptote of zero by the EMIRM. These same four items were estimated to have much higher lower-asymptotes by the standard 3PL model.

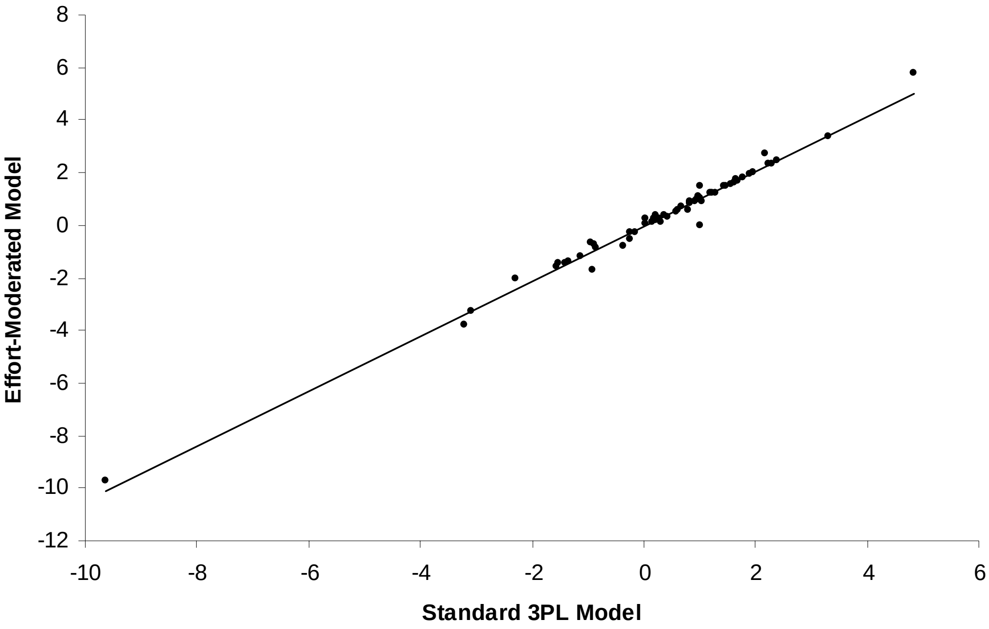

FIGURE 2. *Scatter plot of the item difficulty parameter estimates obtained via the standard 3PL and effort-moderated IRT models.*

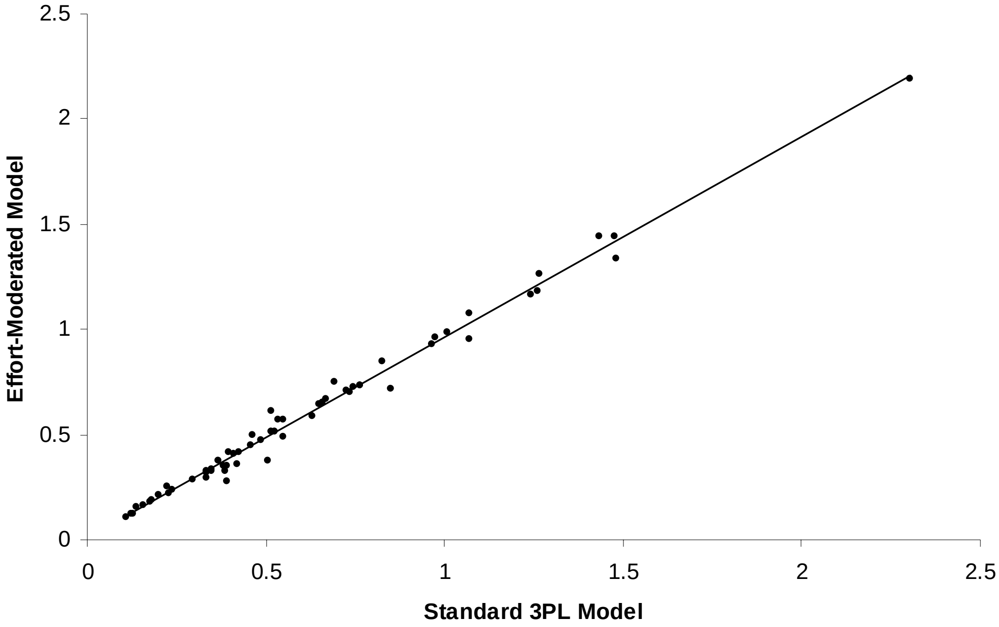

FIGURE 3. *Scatter plot of the item discrimination parameter estimates obtained via the standard 3PL and effort-moderated IRT models.*

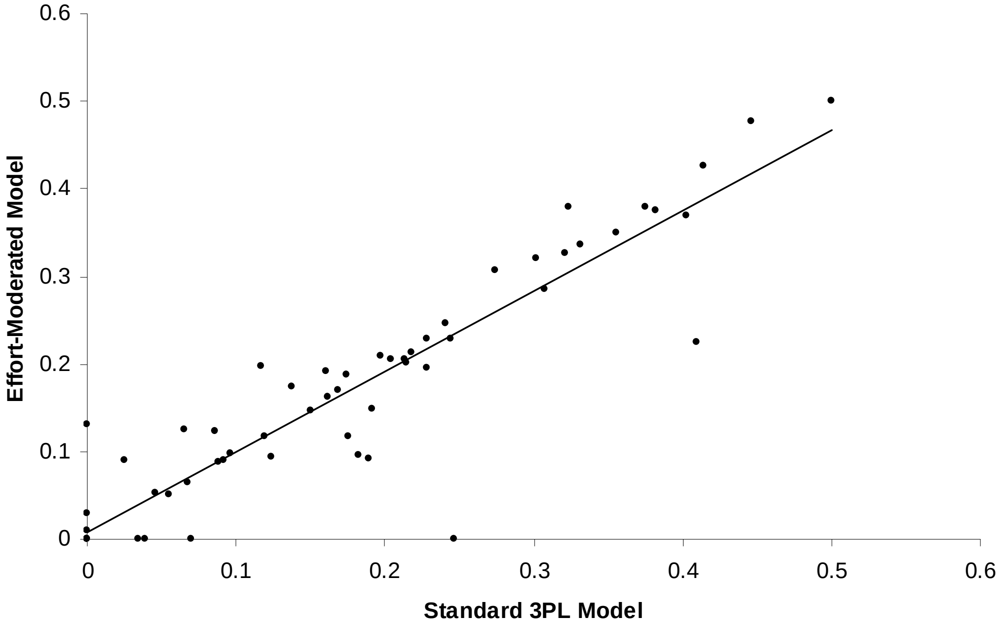

FIGURE 4. *Scatter plot of the item pseudo-guessing parameter estimates obtained via the standard 3PL and effort-moderated IRT models.*

Table 2 provides a slightly different look at the difficulty and discrimination estimates. Here, comparisons are made between the mean estimates at various levels. For example, the discrimination values obtained from the standard 3PL model were sorted from smallest to largest and then separated into three groups (least discriminating to most discriminating). The means of the values in each of the three groups were then calculated. The same process was applied to the discrimination values obtained via the EMIRM for comparison. It appears that greater differences are found between the most discrimination items, at least relative to less discrimination items. However, the mean difference is still quite small. There also appears to be greater differences between the least and most difficult items when compared to the middle third. Note that the difference between the most difficult items is negative (i.e., the EMIRM parameter estimates are higher) whereas the difference for the least difficult is positive.

*Test information.* Further comparisons between the models were made based on the estimated test information function (TIF). Figure 5 displays the TIF for both the EMIRM and standard 3PL model. As can be seen, the TIFs are very similar across the majority of the ability range. The difference seems to occur between values of 0 and just over 2. The peaks of both TIFs are found within this range of ability values as well. This information indicates that the standard errors associated with the theta values in this range will differ across the models. In this case, the standard 3PL model will exhibit smaller standard errors than the EMIRM.

Table 2

*Mean Parameter Estimates*

| | | Mean Parameter Value | | |
|---|---|---|---|---|
| Item Group | Number of Items | Standard 3PL | EMIRM | Mean Difference |
| Discrimination parameter | | | | |
| Least discriminating third | 20 | 0.25 | 0.24 | 0.01 |
| Middle third | 20 | 0.51 | 0.50 | 0.01 |
| Most discriminating third | 20 | 1.08 | 1.04 | 0.04 |
| All items | 60 | 0.61 | 0.59 | 0.02 |
| Difficulty parameter | | | | |
| Least difficult third | 20 | -1.48 | -1.58 | 0.10 |
| Middle third | 20 | 0.65 | 0.60 | 0.06 |
| Most difficult third | 20 | 1.91 | 2.03 | -0.12 |
| All items | 60 | 0.36 | 0.35 | 0.01 |

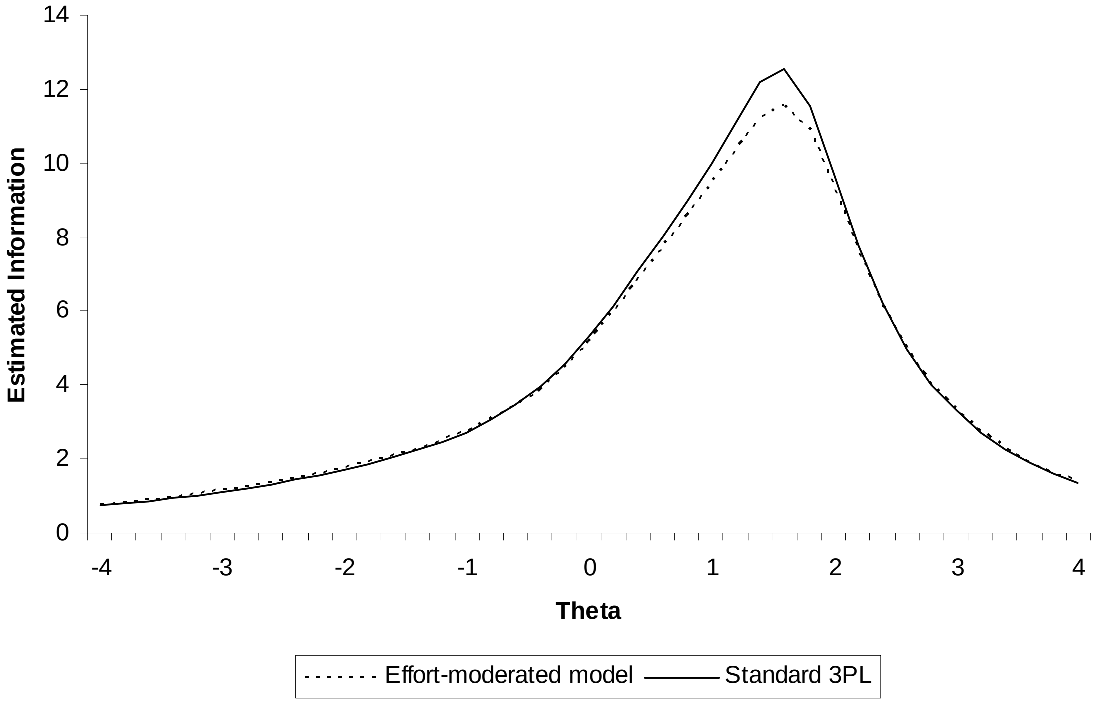

FIGURE 5. *Test information functions for the standard 3PL and effort-moderated IRT models.*

*Convergent validity.* Finally, the theta estimates from both models were correlated with self-reported measures of both overall grade point average (GPA) and GPA within the major. These correlations, along with the correlation between the theta estimates, can be seen in Table 3. In this case, the correlations between thetas from the IRT models and both measures of GPA were virtually identical. Additionally, the theta estimates were very highly correlated.

Table 3

*Correlations among Ability Estimates*

| | Overall GPA | Major GPA | Theta-3PLM | Theta-EMIRM |
|---|---|---|---|---|
| Overall GPA | 1.000 | | | |
| Major GPA | 0.755 | 1.000 | | |
| Theta-3PLM | 0.373 | 0.349 | 1.000 | |
| Theta-EMIRM | 0.374 | 0.351 | 0.997 | 1.000 |

Note: n = 7,494 (excludes any missing data on the GPA measures).

## Discussion

The current study provided an application of the effort-moderated model to a large-scale, low-stakes test and is for the most part a replication of the original study by Wise and DeMars (2006). Unlike the results reported by Wise and DeMars, in our study, very few differences were found in terms of difficulty and discrimination parameter estimates, as well as ability estimates. Some differences were found in the pseudo-guessing estimates, although these appeared to have only a minor effect in the long run.

Small differences were also found in the test information function, where the standard 3PL model tended to have higher information around the peak of the function. As stated earlier, this would lead to smaller standard error estimates for thetas in that range (and different reliabilities). In the current study, we have no way of knowing whether the standard 3PL model overestimated the test information or if the EMIRM underestimated the TIF. However, the results of simulations have shown that the standard 3PL model does in fact underestimate standard errors in the range where test information is greatest. We suggest that this is likely to also be the case with these data.

One difference between the current study and the original work by Wise and DeMars is the number of examinees who exhibited any level of rapid-guessing behavior. In our study, approximately 91 percent of the examinees had RTE scores of 1.0. Wise and DeMars reported that only 69 percent of their sample had RTE scores of 1.0. In their second study using simulated data, one of the conditions contained very similar RTE scores to those exhibited here (i.e., 90 percent with RTE=1). However, those authors still reported rather significant differences in the test information across the majority of theta values. One potential reason for the difference in findings is the fact that 2.3 percent of all responses in their simulated data set (in the 90 percent condition) were rapid-guesses compared to only 1 percent in our data set.

Another reason why fewer differences were found could be the fact that the MFT test used for these analyses is a particularly difficult one. Whereas the mean ability estimates for both models were approximately zero, the peaks of the test information functions were somewhere between positive one and two. It could be argued that a test comprised of easier items would result in greater discrepancies between the model-implied probability of a correct response and the true probability when guessing occurs. For example, consider the two item characteristic curves (ICC) in Figure 6. The ICC for item B indicates that it is more difficult than item A. Note that if a low-ability examinee with a theta of say -1.0, encounters the difficult item he or she has little probability of correctly answering the item. This is the case regardless of whether the random-guessing probability is assumed (i.e., approximately .20) or the standard 3PL model-implied probability is used. Now consider the estimated probabilities for item A. At a theta level of -1.0, there is a much larger discrepancy between the random-guessing probability of .20 and the model-implied probability.

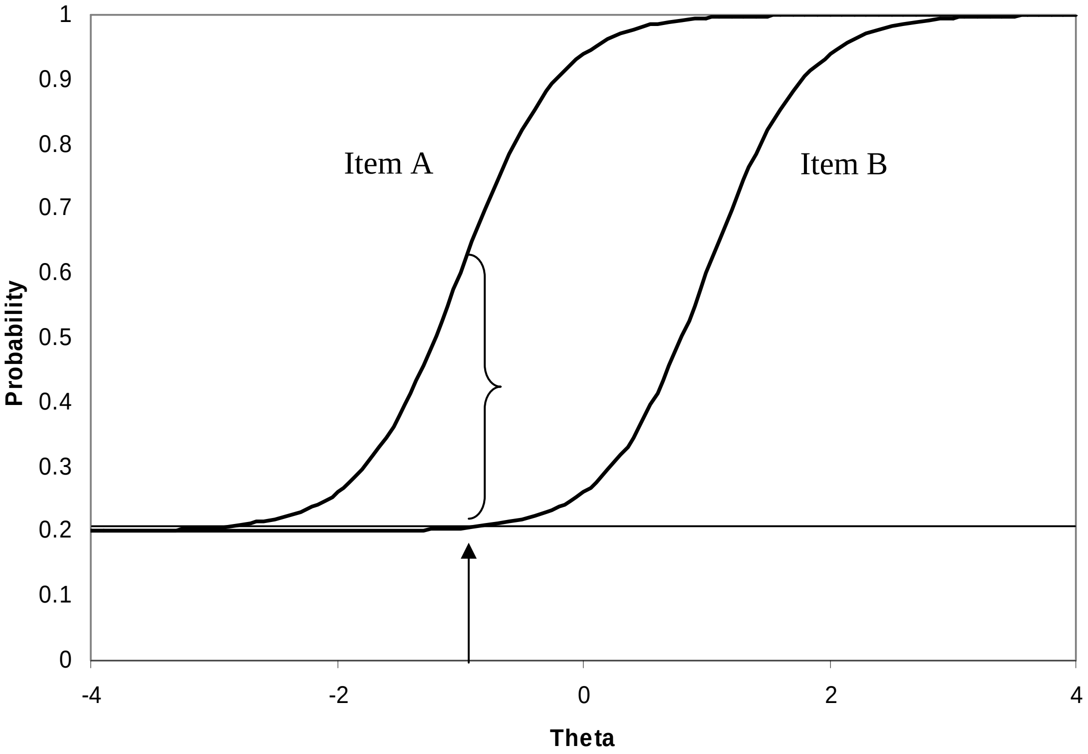

FIGURE 6. *Using item characteristic curves to demonstrate the effect of item difficulty on the discrepancy between the standard 3PLM and the EMIRM.*

Based on the conclusions presented thus far, one may deduce that there is little difference between the standard 3PL IRT model and the EMIRM. However, as we have mentioned, only the first half of the total test was used for these analyses. When we added the second set of 60 items and recalculated the number of examinees with RTE equal to one, the number dropped from 91 percent to 67 percent! Unfortunately, the lack of unidimensionality precluded us from applying an IRT model to the entire test. Nevertheless, this fact suggests that greater differences between the standard 3PL model and the EMIRM may have been found with these data. It also suggests that there is likely a fatigue factor that may affect the overall test results.

There remain many directions in which this research could be expanded. For example, in our study and in others that have used the RTE measure (Wise & DeMars, 2006; Wise & Kong, 2005), solution-behavior has been treated as a dichotomous measure of effort. Whether this is the most appropriate treatment of the measure remains to be seen. Future research could explore other ways of using a continuous version of the measure.

An additional area of research in which response time could be explored involves the cognitive processing of the examinee. Is it possible to use response time to better understand the interaction between the examinee and the test items? Some research has already begun in this area (e.g., Wise, Pastor, & Kong, 2007) but certainly more is needed. With the recent advances in psychometric models that incorporate cognitive processes, this area of research is rich with opportunity.

# References

Kong, X., Wise, S. L., Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution-behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*(4), 606-619.

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. (ERIC Document Reproduction Service No. ED400276)

Schnipke, D. L. (1999). The influence of speededness on item-parameter estimation (Computerized Testing Rep. 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 C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Sundre, D. L. (1999). *Does Examinee Motivation Moderate the Relationship between Test Consequences and Test Performance?* (Report No. TM029964). Harrisonburg, Virginia: James Madison University. (ERIC Document Reproduction Service No. ED432588).

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education, 19*(2), 95-114.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19-38.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1-17.

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program.* Paper presented at the National Council of Measurement in Education annual conference: San Diego, CA.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183.

Wise, S. L., Pastor, D. A., & Kong, X. (2007, April). *Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice.* Paper presented at the National Council of Measurement in Education annual conference: Chicago, IL.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227-242.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3 for Windows.* Lincolnwood, IL: SSI Scientific Software International.