The Alberta Journal of Educational Research *Vol. XLIX, No. 3, Fall 2003, 277-289*

Christina van Barneveld
Lakehead University

The Effects of Examinee Motivation on Multiple-Choice Item Parameter Estimates

The purpose of this study was to examine the effects of false assumption regarding the motivation of examinees on item calibration and test construction. A simulation study was conducted using data generated using two models of item responses (the 3-parameter logistic item response model alone, and in combination with Wise's Examinee Persistence model (1996a)). Items were calibrated using a Bayesian method. Results clearly document the effect of low motivation on item parameter estimates. For the conditions studied, the item parameter estimates based on responses from poorly motivated examinees were biased and more variable than estimates based on responses from examinees who behaved according to the three-parameter logistic model.

Low motivation test-taking behaviors may occur when the examinee is aware that the results from portions of the test (or the complete test) have no personal consequence. This circumstance may arise when the purpose of the test administration is to pilot test items or to establish group-level scores, as in statewide or provincial assessment. A student with low motivation may not persist in applying his or her abilities when responding to test items, opting instead to guess, omit a large number of item responses, or quit entirely. This examinee's response pattern is aberrant because performance on the inconsequential items does not reflect his or her ability.

Earlier research in the area of test-taker motivation suggests that item and test characteristics are related to the motivation of examinees. Stocking, Steffen, and Eignor (2001) studied the responses of more than 33,000 college-level examinees to an operational Graduate Record Examination (Analytical) and found that the proportions of examinees who randomly guess (defined as spending less than 10 seconds per item) or omit the response entirely increase as item position increases (i.e., as the test goes on). They also observed that the proportions of examinees who guess or omit are higher for items that are part of a set of items with a common stimulus than for discrete items. They concluded that item position and item format influence test-taking behaviors. Item format effects were also observed by DeMars (2000), who studied 11,930 high school students writing the Michigan High School Proficiency Test. She found that the average score of students was higher when the stakes were high (diploma endorsement) than when stakes were low (pilot test), but the difference was larger for constructed response than multiple-choice items, suggesting that item format may interact with motivation to perform. Wolf, Smith,

---

Christina van Barneveld is a lecturer in the Faculty of Education. Her current research interests include educational measurement and evaluation, aberrant examinee response behaviors, and reliability and validity issues related to assessment instruments applied during the training and certification of professionals in the fields of education and medicine.

and Birnbaum (1995) considered the responses of 301 high school students to a New Jersey state graduation test in math. They studied item difficulty, task exertion (operationally defined as a rating provided by six experienced math educators), and item position as predictors of differential item functioning (DIF) between the groups of examinees in high- and low-stakes circumstances. They found DIF between the low- and high-stakes groups was predicted by item difficulty and task exertion. The correlation between item difficulty and task exertion was .40. Interestingly, DIF between low- and high-stakes groups was not predicted by item position. They note, however, that item position in this test did not vary and that the test length was only 30 items long, perhaps too short to produce fatigue. Wolf and Smith (1995) observed 158 college students on both a high-stakes and a low-stakes class test and collected self-report information regarding student motivation. Examinees reported higher levels of motivation for the high-stakes test. Also, most examinees had slightly higher scores on the high-stakes test, although one third of them had higher scores on low-stakes test.

Research has also delved into the relationship between examinee characteristics and low motivation test-taking behaviors. Stocking et al. (2001) found that the proportion of examinees who randomly guessed or omitted items increased as ability decreased. This effect was also noted by Wise (1996a, 1996b), and Wolf and Smith (1995). Other examinee characteristics such as gender and ethnicity appear to be less influential in determining test-taking motivation than ability (DeMars, 2000).

Item Response Models

Item response models describe the relationship between an examinee and the items on a test. Recent research has focused on the identification of unlikely response patterns in relation to item response models (Kalohn & Spray, 1999; Nering, 1997; van Krimpen-Stoop & Meijer, 1999). Many types of item response models have been developed (van der Linden & Hambleton, 1997). In this study, attention is focused on a three-parameter logistic item response model for dichotomously scored data.

The Three-Parameter Logistic Item Response Model

The three-parameter logistic (3PL) item response model (IRM) can be defined as

$$P_i(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta_j - b_i)}}, \quad (1)$$

where

- e is the base of the system of natural logarithms,
- i indexes test item (i=1, 2, 3,..., n),
- j indexes examinee (j=1, 2, 3,..., N),
- a_i is the discrimination parameter for item i and is proportional to the slope of the item response function at the point $\theta_i=b_i$,
- b_i is the difficulty parameter for item i, the point on the ability scale at which an examinee has $(1+c)/2$ probability of answering item i correctly,
- c_i is the lower asymptote parameter of the item response function for item i and represents the probability of an examinee with low ability correctly

answering the item (sometimes referred to as the pseudo-guessing parameter),

θ_j represents the ability of examinee *j*,

$P_i(\theta_j)$ is the probability of examinee *j* with ability θ answering item *i* correctly, and

D is a scaling factor; when set to 1.702, the parameter estimates for the logistic and the normal ogive models are comparable.

In order to obtain estimates of the item parameters *a*, *b*, and *c*, an item calibration study is conducted. In general, calibration studies involve identifying a suitable number of examinees, administering the test items, and using the resulting item responses to estimate item parameters. Item parameter estimates are then reviewed, and those that are consistent with the goal of the test are selected for subsequent operational test administrations to obtain ability estimates of future examinees.

It is generally assumed that the group of examinees used to calibrate test items is composed of *normal* responders (Drasgow, Levine, & Williams, 1985; Yi, 1998), that is, they do not display aberrant response behaviors. Although the assumption of a normal calibration group is mentioned in the literature, few studies have focused on the effect of its violation. Violation of this assumption, however, may result in errors in item parameter estimates.

Errors in item parameter estimates may have serious implications for test construction. For example, an overestimation of the item discrimination parameter *a* results in errors in overestimation of the item information (Hambleton & Jones, 1994; Hambleton, Jones, & Rogers, 1993; Tsutakawa & Johnson, 1990; van der Linden & Glas, 2000)—defined as the extent to which the item determines the value of the ability being measured (McDonald, 1999). Information is related to the precision with which ability is estimated, such that the greater the information there is at a given θ, the more precise the measurement will be at θ. It has various applications in the measurement field, such as in describing items, selecting items for test construction, assessing the precision of measurement, and comparing items. In the case of the 3PL IRT model the item information function is expressed,

$$I_i(\theta) = \frac{(1.7)^2 a_i^2 (1 - c_i)}{[c_i + e^{1.7a_i(\theta - b_i)}][1 + e^{-1.7a_i(\theta - b_i)}]^2}, \quad (2)$$

where the terms are defined as in Equation 1. From Equation 2 one can observe the relationship between the item parameters and item information. Information is higher when the *b* value is closer to θ, when the *a* parameter is larger, and as *c* approaches zero (Hambleton, Swaminathan, & Rogers, 1991). Because item information has various applications in the measurement field, opportunities abound to misuse items when there are errors in the item information estimates.

Model of Low Motivation Test-Taking Behavior

Models of low motivation test-taking behaviors have been proposed (Stocking et al., 2001; Wise, 1996a, 1996b). Here I focus on Wise's model because it was the most widely applied at the time of this publication.

Wise's (1996a, 1996b) Examinee Persistence (EP) model is a three-state Markov process—attentive, random guessing, or nonresponse—where all examinees are initially assumed to be in the attentive state. After each item, there is some probability that the examinee will transition from an attentive state to either a random guessing state or a nonresponse state. Examinees in the random guessing state also have some probability of transitioning to the nonresponse state. The nonresponse state is absorbing.

In Wise's (1996a) EP model transition probabilities were modeled as a logistic function of effort, operationally defined as the difference between item difficulty and examinee ability. The transition probabilities were defined

$$r_{ij} = P(R_j = i \mid A_{ij}) = \frac{1}{1 + e^{r_b - r_a(b_i - \theta_j)}}, \tag{3}$$

and

$$q_{ij} = P(Q_j = i \mid Q_j > i - 1) = \frac{1}{1 + e^{q_b - q_a(b_i - \theta_j)}}, \tag{4}$$

where

r_{ij} is the probability that examinee j will transition from the attentive to the random guessing states after item i,

R_j is the random variable denoting the item after which examinee j began to respond randomly,

A_{ij} is a logical variable denoting the event that examinee j responded *attentively* to item i,

r_a is the slope of the logistic function that gives the probability r_{ij},

r_b is the intercept of the logistic function that gives the probability r_{ij},

q_{ij} is the probability that examinee j transitions from either the attentive or the random guessing state to omitting after item i,

Q_j denotes the number of items reached by examinee j,

q_a is the slope of the logistic function that gives the probability q_{ij}, and

q_b is the intercept of the logistic function that gives the probability .

The transition probability matrix described by Wise (1996a) was therefore as shown in Table 1.

In order to provide an interpretation of the various probabilities presented above, consider the following example. For transition probability parameters $r_a = 0.47$, $r_b = 5.54$, $q_a = -0.10$, and $q_b = 5.66$, the proportion of examinees in the attentive state after 50 items may be calculated by $(1 - r_{ij} - q_{ij})^{50}$. For an easy 50-item test (i.e., $b - \theta = -4$ throughout the test), 75% of examinees remain in an attentive state. When the test is difficult (i.e., $b - \theta = 4$), 25% of examinees remain in an attentive state.

To assess this model, Wise (1996a) observed 20,025 army recruits who responded to items (divided into 6 sets) on a low-stakes, paper-and-pencil administration of the United States Army's Computerized Adaptive Screening Test (CAST) word knowledge (WK) and arithmetic reasoning (AR) subtests. Each test form was administered in forward and reverse order to randomly equivalent groups of examinees. He investigated whether the EP model fitted these data better than a model that did not include transitions to random responding, whether transition probabilities varied by item type, by examinee ability, or an interaction of the two. Results indicated that the EP model led to

Table 1
Transition Probability Matrix

		(from)		
attentive	*random-guess*	*nonresponse*		
$1-r_{ij}-q_{ij}$	0	0	*attentive*	(to)
r_{ij}	$1-q_{ij}$	0	*random-guess*	
q_{ij}	q_{ij}	1	*nonresponse*	

significantly improved fit in comparison with the traditional 3PL IRT model. Transition probabilities varied as a function of item type (arithmetic reasoning versus word knowledge), with items that required more effort (i.e., arithmetic reasoning), resulting in higher probabilities of transition. Transition probabilities were higher following more difficult items, especially for low-ability examinees. More detailed accounts of the theory and applications of the EP model are offered by Wise (1996a, 1996b).

Purpose of the Study

The purpose of this study was to provide an example of the potential effect of false assumptions regarding the motivation levels of examinees in low-stakes testing environments on item calibration.

Method

Examinee and Item Characteristics

Simulated ability values of examinees in the calibration group were drawn from an approximately normal distribution with a mean of 0 and a standard deviation of 1.

The item parameters used in this study were those of the 3PL IRT model for dichotomously scored data. For the purposes of this study, the item parameter values used to generate response vectors for simulated examinees were constant. The item parameter values for discrimination, difficulty, and pseudo-guessing were $a = 1.5$, $b = 0$, and $c = .2$ respectively. Although it is unlikely that real-life test items would have identical characteristics, this method was used to isolate the variability introduced by low motivation test-taking behaviors. Any changes in the item parameter estimates were clearly attributable to manipulation of the independent variables.

Design: 3x2x50 with 100 Replications

Three levels of examinee motivation. Three levels of examinee motivation were examined in this study. The first level was defined by attentive examinees who responded to all items according the 3PL IRT model.

The second level of examinee motivation reflected low-motivation test-taker behavior based on the EP model proposed by Wise (1996a), with transition probability parameters $r_a = 0.47$, $r_b = 5.54$, $q_a = -0.10$, and $q_b = 5.66$. Note that these values were the mean of the estimated transition probability parameters based on empirical research on a fixed-item, paper-and-pencil administration of the CAST/AR test (Wise, 1996a). These transition probability parameters were constant over items.

The third level of examinee motivation reflected very low-motivation test-taker behavior based on the EP model proposed by Wise (1996a), with transi-

tion probability parameters $r_a = 0.59$, $r_b = 5.04$, $q_a = -0.15$, and $q_b = 5.54$. Note that these values were the largest absolute values of the slopes from the estimated transition probability parameters based on empirical research by Wise. These transition probability parameters were constant over items.

It was assumed that the artificial test was not speeded. This was done to isolate the effect of examinee persistence, with transition probabilities that were dependent on the difference between the examinee's ability and the difficulty of the item, and not on time taken or time remaining.

Two levels of sample size. The number of simulated examinees used in this investigation were n=500 and n=2,000. The n=500 condition reflected the minimum number of examinees required for calibration of items for a 3PL model (Hulin, Lissak, & Drasgow, 1982). The n=2,000 condition is generally considered to be a large sample from which to obtain satisfactory item parameter estimates (Lord, 1980) and has been used in studies that have examined the influence of errors in item parameter estimates on item information functions and test construction (Hambleton & Jones, 1994; Hambleton et al., 1993; van der Linden & Glas, 2000).

Fifty item positions. There were 50 item positions in the artificial test. This test length was selected to simulate the number of items an examinee may respond to during a calibration study. Items with identical characteristics were administered to simulated examinees in every position in the 50-item test.

Replications. In this design there was the potential for sampling error associated with the sampling of examinees. Therefore, there were 100 samples of examinees from a population for each condition. This resulted in 100 within-condition observations.

Procedure

Data Generation Method

Item responses for examinees were generated using the following procedure. All examinees started the test in an attentive state.

Step 1. Generate an item response for an attentive examinee. To do this, the item response was coded as 1 (correct) if a random number, r, drawn from U(0,1) is less than or equal to $P(\theta)$, and as 0 (incorrect) if $r>P(\theta)$.

Step 2. Determine if the examinee transitions to a random guessing state. To do this, draw a random number from U(0,1) and determine if it is larger than r_{ig} defined in Equation 1. If the random number is larger than r_{ig}, then the examinee does not transition to a random state. Go to Step 3. If the random number is less than r_{ig}, then the examinee transitions to a random state. Go to Step 4.

Step 3. Determine if the examinee transitions to a nonresponse state. To do this, draw a random number from U(0,1) and determine if it is larger than q_{ij} as defined in Equation 8. If the random number is larger than q_{ij}, then the examinee does not transition to a nonresponse state. If the random number is less than q_{ij}, then the examinee does transition to a nonresponse state.

Step 4. Administer next item.

Step 5. If the examinee was in an attentive state, go to Step 1. If the examinee was in a random guessing state, go to Step 6. If the examinee was in a nonresponse state, go to Step 7.

Step 6. Determine if the examinee randomly guessed the correct answer. To do this, draw a random number from U(0,1) and determine if it is larger than c_i. If the random number is larger than c_i, then the item score is 0, reflecting an incorrect response. If the random number is equal to or less than c_i, then the item score is 1, reflecting a correct response. Go to Step 3.

Step 7. The item score is 0 reflecting an incorrect response. Go to Step 4.

Item Parameter Estimation Method

Item parameter estimates for a 3PL IRM were obtained using a marginal maximum likelihood estimation method with an expectation/maximization algorithm (MMLE/EM) approach with Bayesian priors on item parameters using the software BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). Item priors were the default values for the BILOG-MG software as follows: a was distributed as lognormal with a mean of 1.13 and a standard deviation of 0.6 (Baker, 1992), b was distributed as normal with a mean of 0 and standard deviation of 2, and c was distributed as beta with parameters ALPHA=5 and BETA=17 (Swaminathan & Gifford, 1986).

Analysis

Item parameter estimates were averaged over replications and were examined graphically using scatterplots, as a function of sample size, examinee motivation level, and item position.

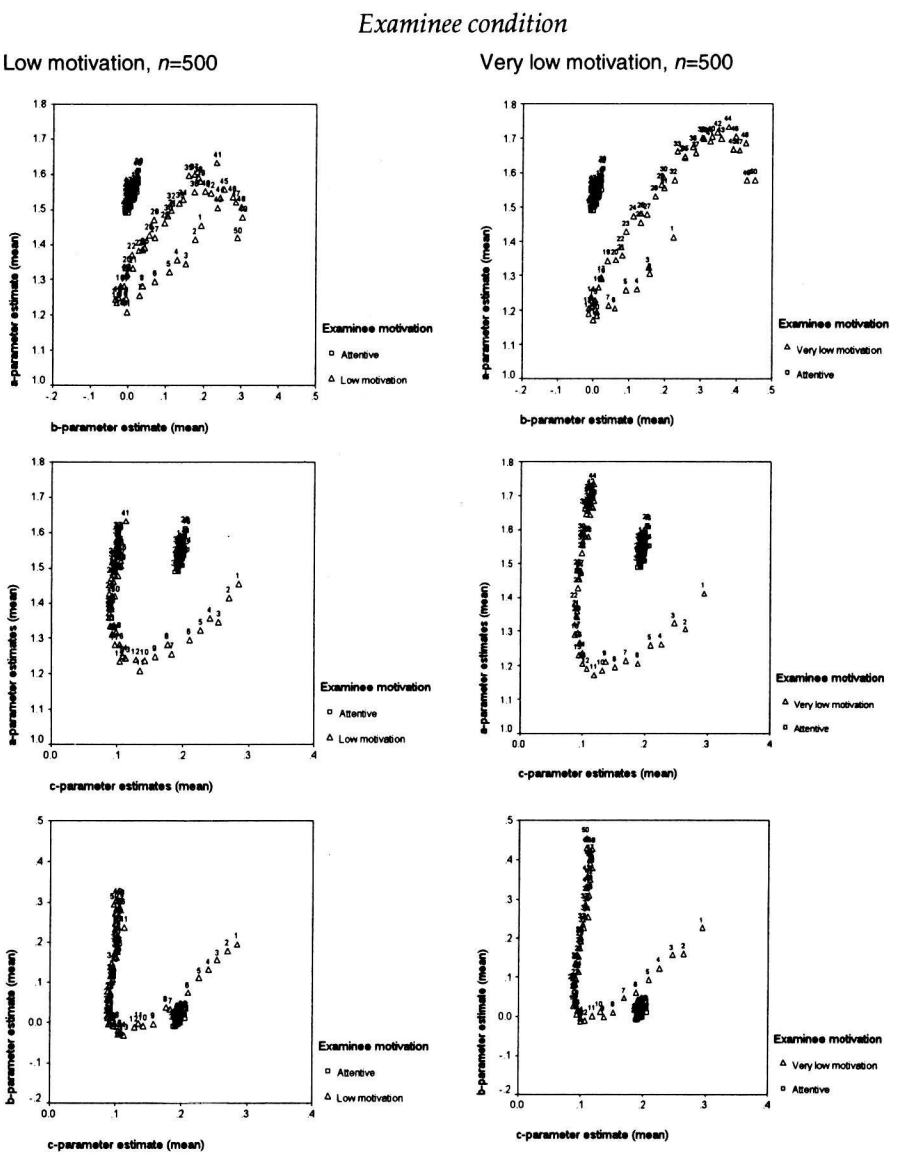
Results

Scatterplots of the mean item parameter estimates (over 100 replications) for an item with true parameters a = 1.5, b = 0 and c = .2 are depicted in Figure 1 (n=500) and Figure 2 (n=2,000). Mean item parameter estimates were plotted by examinee motivation and item position.

In the *attentive* examinee calibration condition, item parameter estimates clustered around their true values, with the exception of the item discrimination parameter. Item discrimination tended to be overestimated (maximum bias of about 0.13) in the *attentive* examinee condition when the calibration group size was small n=500). This effect was not observed when calibration group size was 2,000 examinees.

Item parameter estimates based on item responses from examinees in the low-motivation calibration conditions were biased and more variable than those based on the responses of examinees from the attentive group. Item discrimination parameters tended to be underestimated for the low motivation conditions (by as much as 0.4), for items in positions 1 to 30 for both sample sizes. After the 30th item position in the test, the a parameter tended to be overestimated for n=500 (by as much as 0.25), but remained underestimated for n=2,000. Further, the variance in the mean a parameter estimates were influenced by examinee low motivation. Although the standard error of the mean a-estimates appeared to be smaller for the n=2,000 condition than the n=500 condition, both low motivation conditions tended to have more variation in the a estimates compared with the attentive examinee calibration condition.

The b parameter estimates also appeared to be biased and more variable as a result of the low motivation of examinees. In general, item difficulty paramet-

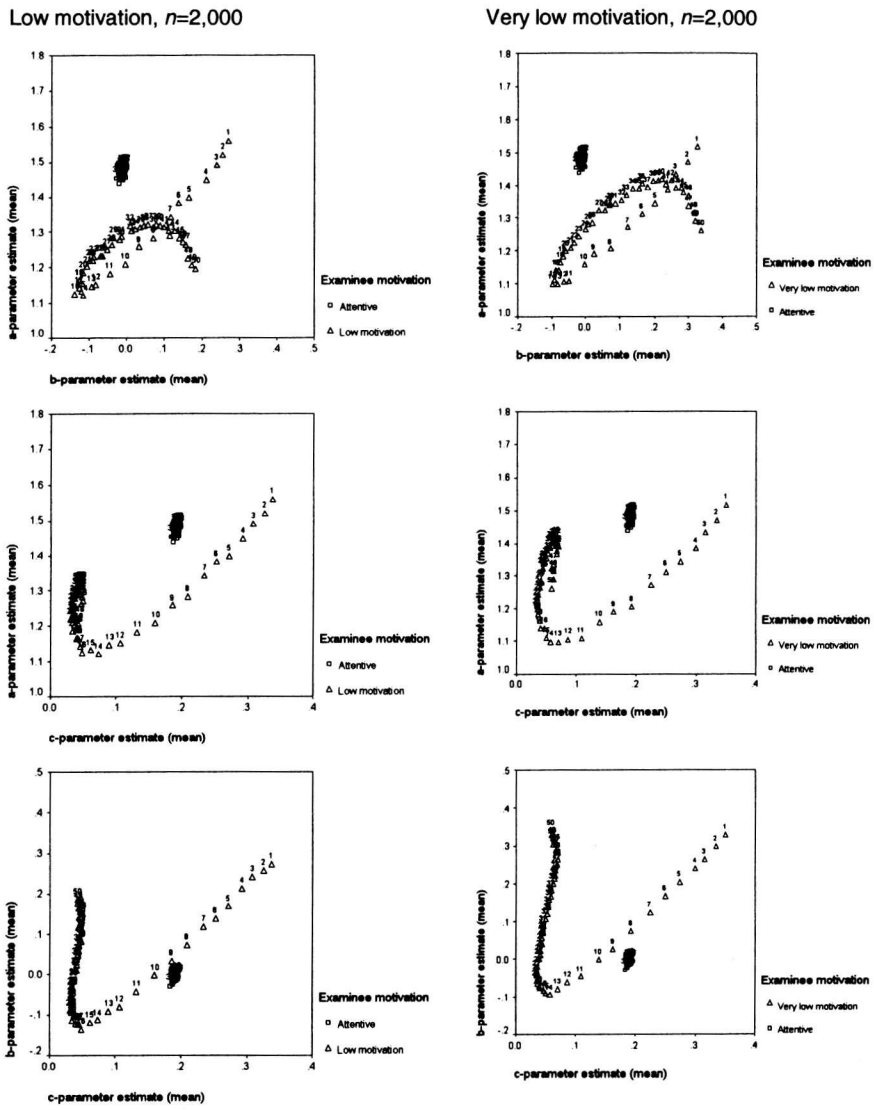


Note. The numbers in each graph represent the item position associated with the adjacent symbol.

Figure 1. Scatterplots of mean item parameter estimates (over 100 replications) for an item with true parameters a = 1.5, b = 0 and c = .2, calibration group size n=500, by examinee motivation and item position. Note. The numbers in each graph represent the item position of

ers tended to be overestimated for both low motivation conditions (by as much as 0.45) as item position increased, especially for the small sample size.

There was a slight overestimation of the pseudo-guessing parameters for items that were positioned early in the test (positions 1 to 8, by as much as .1) and underestimation for items that were positioned toward the end of the test



Note. The numbers in each graph represent the item position associated with the adjacent symbol.

Figure 2. Scatterplots of mean item parameter estimates (over 100 replications) for an item with true parameters a = 1.5, b = 0 and c = .2, calibration group size n=2,000, by examinee motivation and item position.

(positions 9 to 50, by as much as .15) for the very low motivation calibration condition).

For the low motivation calibration conditions, item parameter estimates appeared to be correlated to each other in a nonlinear relationship.

Discussion

Low motivation of examinees in a calibration study results in errors in item parameter estimates, especially when the calibration group is small (n=500). Although errors are smaller for the larger calibration group size, the item parameter estimates are not recovered to their true values.

An interesting finding of this study was the correlation between the item parameter estimates as a result of low examinee motivation. This may be a result of poor fit of the 3PL IRT model to the low motivation examinee responses. This also may be related to the potential introduction of local item dependencies as a result of the random guessing and omitting behaviors displayed by low motivation examinees, especially toward the end of the test where random guessing or quitting was more prevalent. In the low motivation conditions, examinee abilities as specified by the 3PL IRT model were not the only factors influencing examinees' responses to test items. Earlier research has shown that violations to the IRT assumption of local item independence result in bias in item difficulty estimates and item discrimination estimates, overestimation of the precision of examinee scores, and overestimation of test reliability and test information (Oshima, 1994; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Yen, 1993; Zenisky, Hambleton, & Sireci, 2002). This may lead to inaccurate inferences regarding examinee ability that may result in a higher chance of misclassification when making pass/fail decisions based on test results.

Another interesting finding of this study was that after the 30th item position in the test, the *a* parameter was overestimated for n=500 for the very low motivation condition, but remained underestimated for n=2,000. This effect may have been caused by items at the end of the test having few (if any) correct responses, because the transition probability parameters used in the *very low motivation* condition represented extreme values based on empirical research. As a result, the Bayesian approach to the estimation of item parameters may have influenced the values toward the end of the test because a prior distribution was defined for the *a* parameter. The contribution of the prior distribution of *a* depended on the extremity of the estimated value. Greater shrinkage occurred when the estimate and the mean of the prior distribution were substantially different. The prior distribution tended to restrain estimates from assuming unlikely values, as defined by the prior. This effect was reduced for the larger sample size, where there was a higher likelihood of examinees with high ability being included in the sample who remained attentive throughout the test.

Errors in the item parameter estimates have implications for test development organizations. For example, when the discrimination parameter is underestimated, an item appears to be less informative than it truly is at a given ability level. Depending on the purpose of the test and the item selection criteria (e.g., maximum information selection criteria), this item may be underused or dropped entirely. When several items are affected in this way, the cost of developing and testing items is increased, and field-test results may not accurately describe the appropriateness of the item given the purpose of the test. For the testing organization responsible for the test, underestimation of

item discrimination may result in an inefficient use of resources required to construct and administer the test items.

The results of this study also have implications for future test-takers. Items that appear to be more difficult than they really are may be misused when constructing an operational test. For example, when items with overestimated difficulty parameters are administered to a high-motivation group of examines—as may occur in a high-stakes operational test administration—their ability estimates will be overestimated because the examinees appeared to answer difficult questions correctly. In fact, these items are less difficult than they appear.

Low motivation response behaviors in the calibration group may also have implications for the development of a computerized adaptive test (CAT). For example, in a CAT where a maximum information strategy is employed as the item selection algorithm, items with underestimated information may not be selected because they do not appear to improve precision of the ability estimate. Finally, errors in item parameter estimates can cause errors in the location of the item information curve, shifting it to the right. Each of these errors in the item information function may result in an increase in the error associated with examinee ability estimates, which may require the examinee to respond to more questions than necessary in order to achieve a given standard error of the ability estimate.

Conclusion

This study illustrates the potential effect of false assumptions regarding the measurement model used to describe the relationship between examinees and test items. When data that reflected low motivation test-taking behaviors were calibrated using the 3PL IRM, item parameter estimates were biased and more variable than when the calibration group was composed of *normal* (attentive) examinees. The direction and magnitude of the biases depended on degree of motivation, calibration group size, and item position within the test.

A systematic review of the effect of examinee motivation on item calibration (and test construction) is needed. Potential veins of research include: the relationship between examinee motivation and item calibration for commonly used item response models, the development of reliable tools for the identification of low motivation examinees in a calibration group (perhaps making use of the item position effect within a test), and robust estimation of item parameter estimates.

Limitations of the Study

Items with identical parameters were used in this study. Thus the generalizability of these results is restricted to an item pool with similar characteristics. Other item parameters typically found in large-scale assessments that employ the 3PL IRT model need to be considered. Item parameter estimation errors should be explored as a function of selected a and b parameters. In a future study, it may be interesting to explore the effect of alternate positioning of a given item within a test of items with varying characteristics. Perhaps this may serve as a starting point for the development of an index of low examinee motivation.

The difference between the probability of a correct answer given a random guessing examinee state in Wise's (1996a) model and the *c* parameter defined in the 3PL IRM is unclear. Usually the *c* parameter is very near the inverse of the number of options, suggesting random guessing behavior of the examinee. It may be less than the inverse of the number of options, however, when some distracters are particularly attractive to low-ability examinees. It may also be higher in value than the inverse of the number of options when some distracters can be dismissed by low-ability examinees. Thus the *c* parameter assumes an attentive state, where low-ability examinees are engaged with the item. Only when low-ability examinees are attentive can the *c* parameter reflect test-taker behavior as described above. When the low-ability examinee is in a random guessing state, he or she is no longer engaged with the test items and therefore his or her probability of a correct answer is the inverse of the number of options.

In this study, omitted data were scored as incorrect. This is the harshest possible treatment of missing data and perhaps resulted in more extreme item parameter estimates than if another treatment had been used. In this case, the accuracy of the item parameter estimates decreased as the number of omissions increased. Other treatments of omissions (De Ayala, Plake, & Impara, 2001) may be considered in a future study.

References

De Ayala, R.J., Plake, B.S., & Impara, J.C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38,* 213-234.

DeMars, C. (2000). Test stakes and item format interaction. *Applied Measurement in Education, 13,* 55-77.

Baker, F.B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

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

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

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

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hulin, C., Lissak, R., & Drasgow, F. (1982). Recovery of 2-parameter and 3-parameter logistic item characteristic curves—A Monte-Carlo study. *Applied Psychological Measurement, 6,* 249-260.

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

Lord, F. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

McDonald, R. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

Nering, M. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21,* 115-127.

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

Sireci, S.G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237-247.

Stocking, M., Steffen, M., & Eignor, D. (2001). *A method for building a realistic model of test taker behavior for computerized adaptive testing* (Draft research report). Princeton, NJ: Educational Testing Service.

Swaminathan, H., & Gifford, J. (1986). Bayesian-estimation in the 3-parameter logistic model. *Psychometrika, 51,* 589-601.

Thissen, D., Steinberg, L., & Mooney, J. (1989) Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26,* 247-60.

Tsutakawa, R., & Johnson, J. (1990). The effect of uncertainty of item parameter-estimation on ability estimates. *Psychometrika, 55,* 371-390.

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

van der Linden, W., & Hambleton, R. (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

van Krimpen-Stoop, E., & Meijer, R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23,* 327-345.

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

Wise, L.L. (1996b). *Indicators of student effort on the National Assessment of Educational Progress.* (Tech. Rep. for Grant No. R99B60002).

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

Wolf, L.F., Smith, J.F., & Birnbaum, M.E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8,* 341-351.

Yen, W.M. (1993). Scaling performance assessments—Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213.

Yi, Q. (1998). *A comparison of three ability estimation procedures for computerized adaptive testing in the presence of nonmodel-fitting responses resulting from a compromised item pool.* Unpublished doctoral dissertation, University of South Florida, Tampa.

Zenisky, A.L., Hambleton, R.K., & Sireci, S.G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement, 39,* 291-309.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, R. (1996). *BILOG-MG.* Chicago, IL: Scientific Software International, Inc.