# An IRT-based response likelihood approach for addressing test speededness

**Andrew A. Mroch**

**Daniel M. Bolt**

**University of Wisconsin-Madison**

**Paper presented at the 2006 annual meeting of the National Council on Measurement in Education, San Francisco, CA.**

## Abstract

An item response theory (IRT) based procedure is proposed for reducing the biasing effects of speededness on end-of-test item characteristics. The procedure characterizes speeded responses as atypical responses (e.g., due to random guesses or omits) under the nominal response model to items at the end of the test. Examinee responses are probabilistically deleted from the calibration data set (i.e., treated as "missing") in proportion to their likelihood of being speeded. Real data from a university mathematics placement test involving common items administered at different locations on different test forms are used to demonstrate implementation of the procedure and evaluate it effectiveness. Application of the procedure is shown to lead to a reduction in the bias of estimated item and test characteristics.

# An IRT-based response likelihood approach for addressing test speededness

## Introduction

Test speededness effects are often observed when examinees do not have sufficient time to finish a test. When examinees are rushed or run out of time, they often fail to adequately answer items at the end of the test. Speededness creates methodological problems in item response theory (IRT), including over-estimates of item difficulty parameters (Bolt, Cohen, & Wollack, 2002; Kingston & Dorans, 1984; Oshima, 1994) and multidimensionality (Lord, 1956; Lord & Novick, 1968; Myers, 1952).

Several methods have now been proposed that model speededness from a single test administration and are capable of reducing the biasing effects of speededness on item and test characteristics. These include the mixture Rasch model (MRM; Bolt, et al., 2002), the multiclass mixture Rasch model (MMRM; Mroch, Bolt, & Wollack, 2005), and the HYBRID model (Yamamoto, 1987; Yamamoto & Everson, 1997). These IRT models address speededness through the introduction of latent examinee classes that are distinguished by speededness effects. Although these models offer interesting possibilities for identifying and reducing the effects of speededness, they have some disadvantages that may limit their utility in practice.

First, these methods assume that examinees move sequentially through the test in answering items; once an examinee is speeded on one item, he/she is assumed to be speeded on all subsequent items. While this assumption may be appropriate for some examinees, others may proceed in a nonsequential fashion. For example, some examinees may solve items they know first, regardless of their location on the test. Second, the methods consider only certain sources of evidence of speededness. For example, the HYBRID model, MRM, and MMRM only consider the correctness of response in modeling speededness. Thus, one of the standard indicators of

speededness often used by testing programs, omitted responses at the end of the test (or, more strictly, not reached items), is not incorporated. Omitted responses at the end of the test are likely to be stronger indicators of speededness than are incorrect responses, making it desirable to retain this information when modeling speededness. On multiple choice items in particular, there may also be value in attending to the specific distractors chosen on incorrect responses, as distractor selection may be related to ability. Consequently, speededness behavior that results in random guessing may also be reflected by the types of distractors selected (e.g., higher ability examinees are more likely to select a distractor representing low ability when guessing randomly).

This study is driven in part by the desire to explore inclusion of end-of-test omits and aberrant end-of-test distractor selection as evidence of speededness. We propose an IRT-based procedure for reducing the effects of speededness that uses the likelihood of item responses as evidence of speededness. Because atypical responses can take several forms (e.g., omits, or unlikely distractor selection), the proposed procedure can account for speededness that emerges in the form of guessing, rushed, and omitted responses.

Compared to existing IRT-based methods for reducing the effects of speededness, the proposed procedure has several anticipated advantages. First, as noted, it uses information contained in the response options as opposed to only the correctness of the responses. Second, it incorporates omitted responses, which provide perhaps the strongest evidence of speededness. Third, it does not assume a sequential ordering in the responses to end-of-test items, and thus relaxes a primary assumption underlying methods such as the HYBRID model, MRM, and MMRM. Fourth, the proposed procedure can be applied to data sets possessing large numbers of

items and examinee responses, without facing the computational limitations of alternative procedures.

The proposed procedure integrates several methodologies, which will be described in the next section. First, we will briefly discuss the nominal response model (NRM; Bock, 1972), followed by a description of a method for incorporating omitted responses. Then we will describe the idea behind the randomized response methodology used in survey research, and also discuss the concept of person fit, two related strategies to that advocated here.

**The Nominal Response Model**

The NRM is a polytomous IRT model that can be used to model the probability of responding in each category of a multiple-choice item. The probability of examinee *j* choosing category *h* for item *i* is modeled by a multivariate logit (Bock, 1972):

$$P(u_{ij} = h \mid q_j, c_{ih}, a_{ih}) = \frac{e^{c_{ih} + a_{ih} q_j}}{\sum_{h=1}^{m_i} e^{c_{ih} + a_{ih} q_j}} \qquad (1)$$

where

$u_{ij}$ is the response category chosen by examinee *j* to item *i* (where categories are $h = 1, 2, \ldots m_i$),

$q_j$ is an ability parameter for person *j*,

$c_{ih}$, is an intercept parameter for category *h* of item *i*,

$a_{ih}$, is a slope parameter for category *h* of item *i*, where

the category intercepts and slopes sum to zero for each item.

Under the NRM, each response category has associated with it a propensity for selection at a given ability level. Implicit in the model is the requirement that, at a given ability level, the probability of the response categories sum to 1; that is, $\sum_{h=1}^{m_i} P(u_{ij} = h \mid q_j, c_{ih}, a_{ih}) = 1$. An example

set of NRM curves for a five-category multiple choice item (plus an omit category) is presented in Figure 1. Each curve illustrates an option response function (ORF), which when portrayed graphically is known as an option characteristic curve (OCC). The OCCs and corresponding ORFs represent the probability that an examinee at a given ability will respond in each response category; this probability is referred to as the response probability. For example, in Figure 1 the ORF with the largest response probability for an examinee with $\theta = 0$ is "C" (the correct response; which occurs with probability .8). Including omitted responses as evidence of speededness is described next.

**Omitted Responses as Evidence of Test Speededness**

Omitted responses are generally good indicators of speededness. However, it is typically not realistic to use omits as the sole indicators of speededness, especially when examinees are encouraged to make guesses when they do not know an answer or run short on time. For a particular test, the utility of using omits as evidence of speededness requires that some amount of omitting behavior be observed. Before including omitted responses as evidence of speededness, we should examine the extent to which omits appear useful to include as evidence of speededness. For purposes of this study, more frequent omitted responses at the end of the test were used to justify including omits as evidence of speededness. As part of the proposed procedure for minimizing the effects of speededness, omitted responses were treated as a separate response category under the NRM. Next, we describe the notion of randomized response, a methodology that inspired the likelihood-based random elimination of responses under the proposed procedure.

**Using Randomized Responses in Testing**

The method of randomized response was originally formulated as a methodology in survey research for obtaining answers to sensitive questions from respondents. These sensitive questions potentially lead to refusal to respond or to giving dishonest answers (Warner, 1965). Randomized response is meant to increase the cooperation of survey respondents by making their responses to a survey truthful with a certain probability. For example, to determine base rates of test cheating, survey respondents would flip a coin (the result of which was only apparent to the respondent). If the coin were heads they would answer "yes", and if it were tails they would answer truthfully. The probability of a random response is known and can be accounted for in marginal estimates of the truthful responses to the sensitive question (Fox, 2005; Warner, 1965).

The general idea behind randomized response is that observed responses have an additional component contributing to them (e.g., social desirability) besides the underlying trait intended to be assessed, and that this additional component leads to aberrant responses. Randomized responses can be used to statistically account for the aberrant responding so as to produce purified estimates of the underlying characteristics of items. Because the goal in an item analysis is to obtain marginal estimates of item characteristics (e.g., IRT estimates of discrimination and/or difficulty), the same general idea can be applied. We seek to determine the likelihood that an examinee response is due to speededness, and in turn reduce the effects of speededness by randomly eliminating examinee responses to end-of-test items in proportion to their likelihood of being speeded. In this way, an "unspeeded data set" can be created that then becomes the basis for calibrating an IRT model. We consider next the concept of person fit,

which similarly employs the use of response likelihoods in evaluating the validity of item responses.

**Person Fit**

A second related methodology to that introduced in this study is appropriateness measurement, also referred to as person fit. Person fit analysis involves identifying examinee response patterns that are unusual or unexpected, given a particular model (e.g., IRT model). Person-fit methods are often based on the assumption of a parametric IRT model (Meijer, 1996, 2003).

Such methods involve a two-step process. First, an IRT model is fit to a sample of normal or typical examinees. Second, a statistic is computed for each examinee to account for the extent to which his/her responses are consistent with the IRT model used to characterize performance (Drasgow, Levine, & Williams, 1985). Such approaches can be subdivided into likelihood-based methods and residual-based methods (Embretson & Reise, 2000; Meijer & Sijtsma, 2001). Likelihood-based methods quantify the likelihood of a response pattern assuming a particular IRT model. Patterns that are not very likely given an examinee's ability are indicative of aberrant responses (Drasgow, et al., 1985). Residual-based methods compare an examinee's observed item response to the item response predicted by the IRT model. Differences between observed and predicted responses quantify aberrant responses (Bejar, 1985; Tatsuoka, 1996). As conceptualized in this study, speededness is expected to lead to unusual patterns of response, and is thus consistent with the general framework of person fit.

Bejar (1985) used a parametric residual-based person-fit statistic to examine simulated data against data from the Test of English as a Foreign Language (TOEFL) for speededness. Bejar found that his person-fit statistic was able to identify the presence of speededness in the

data sets he examined. Unlike Bejar, we use a likelihood based statistic for evaluation of speeded responses, and study speededness on a response by response basis.

**An IRT-Based Randomization Procedure for Addressing Test Speededness**

The IRT-based randomization procedure proposed in this paper for reducing the effects of speededness essentially involves constructing a data set by randomly eliminating examinee responses to end-of-test items from a speeded data set in proportion to their likelihood of being speeded. Examinee responses are probabilistically excluded (i.e., treated as missing) according to their observed response probabilities (i.e., likelihoods) under the NRM. Typical large-scale testing programs test a larger number of examinees than is required for adequate IRT parameter estimation, making the exclusion of examinees from the calibration data set often of negligible cost. However, the loss of examinee data should be weighed against the potential reduction in bias by reducing speededness. As with any method for addressing speededness, its potential utility should be considered for a given test.

The proposed procedure is inspired by likelihood-based person-fit methods but differs from typical methods in that it considers the likelihood of each item response separately, whereas likelihood-based person-fit considers the entire response pattern. A step-by-step description of the procedure follows.

*Step 1. Identify a subset of items that likely contain bias due to speededness effects.* Identification of end-of-test items can be assessed a number of ways. For example, factor analysis can be used to identify items that tap a common secondary speededness factor. Or an examination of omit frequencies can be used to identify which end-of-test items appear to be biased due to speededness. Finally, a survey of examinees could be taken to obtain direct information as to which of their responses were speeded. In our current analysis, we examined

nonlinear factor analysis solutions and omit frequencies to identify items likely to contain bias due to speededness.

*Step 2. Fit a nominal response model (NRM).* Fit the NRM to all items on the test, including omitted responses as a separate response category for all items except the end-of-test items identified in Step 1. The omit category for end-of-test items will be treated differently as described below.

*Step 3. Obtain an estimate of examinee ability uncontaminated by speededness.* This estimate of ability is obtained to reduce the bias that speededness may have on ability estimates. For example, examinee ability estimates could be based on items designated in Step 1 as being unspeeded items.

Obtaining such an estimate of ability may not always be feasible. For example, if 9 items on a 10-item test show evidence of speededness, estimating examinee ability based on one item would naturally be problematic. However, in most applications it is anticipated that the majority of test items will be unspeeded and thus can be used to obtain ability estimates.

*Step 4. Construct an average option response function (ORF) for the omit category across unspeeded items.* This ORF will later be imposed as the omit ORF for end-of-test items. The purpose of constructing this omit ORF independent of the end-of-test items is that speededness should lead to more frequent occurrences of omits, while in an unspeeded situation omits are less likely.

*Step 5. Apply the average unspeeded omit ORF to the option characteristic curves (OCCs) of end-of-test items.* This approximates what the expected omitted response probabilities should look like for end-of-test items when speededness is not present. Using the set of item parameter estimates from Step 2, the constructed omit ORF can still be applied to each end-of-

test item such that, for a given examinee, all response category probabilities for the item will still sum to 1.

*Step 6. For end-of-test items, compute the NRM response likelihood of each observed response.* The original NRM response probability, $P_{ih}(q_j)$, now represents the conditional likelihood that an examinee with ability $q_j$ chooses response category *h* to item *i* from Equation 1, assuming the response is not an omit. This likelihood is based on the updated OCCs for end-of-test items obtained in Step 5 that include the omit category as a possible response. For example, using the OCCs in Figure 1, an examinee with ability -1 that selected category 'E' has a response likelihood of .17.

*Step 7. For end-of-test items, simulate a model-based set of examinee responses and compute the NRM response likelihood of each model-based response.* Using the updated OCCs from Step 5 and examinee ability estimates obtained in Step 3, simulate responses to the end-of-test items and then compute response likelihoods, $P_{ih}(q_j)$, based on these responses. This results in unspeeded model-based simulated item responses for the end-of-test items. For example using the OCCs in Figure 1, if we have an examinee with ability -1 whose simulated response is "C", that person's response likelihood is .62. The reason for simulating this data set is to generate a distribution of response likelihoods consistent with unspeeded responding. These data are used as a comparison against the response likelihoods of the real data from Step 6.

*Step 8. For each end-of-test item, segment the response likelihoods into bins.* Segment the response likelihood values for all examinees into bins separated by intervals (e.g., 0 to .1; .1 to .2; .2 to .3, etc.). This produces a distribution of real response likelihoods and a distribution of simulated response likelihoods for each end-of-test item (see Figure 2a for an example of these distributions).

*Step 9. For each end-of-test item, compute the ratio of real to simulated response likelihoods for each bin.* Compare the real and simulated distributions of response likelihoods for each end-of-test item by computing the ratio of real to simulated frequencies in each bin. This ratio is computed by dividing the frequencies of the real distribution by the frequencies of the simulated distribution.

*Step 10. For each end-of-test item, set a minimum ratio for the real to simulated response likelihoods (to ensure adequate bin sizes) and randomly eliminate item responses in proportion to this ratio.* Compare this minimum ratio to the smallest ratio obtained across bins (computed in Step 9). The smallest bin ratio should be equal to or larger than the minimum ratio. Redefine bins as needed by increasing bin widths (repeat Steps 8 and 9) to obtain a smallest bin ratio equal to or larger than the minimum ratio. To ensure adequate numbers of examinees in each bin, it may be necessary to specify a minimum number of examinees for each bin.

Finally, randomly eliminate observed item responses in proportion to the smallest bin ratio for each item. In bins having ratios greater than the smallest ratio, randomly eliminate (treat as missing) data in each bin to get the bin ratio to the size of the smallest bin ratio, such that the relative distributions of real and simulated response likelihoods are the same. This results in a "purified" data set in which the real and simulation-based response likelihoods are approximately the same. This new data set can then be used for calibrating items via the IRT model used to calibrate the operational test (e.g., Rasch model or 3-parameter model).

*Summary.* Under the proposed procedure, the real response likelihoods quantify the likelihoods of the observed responses to end-of-test items. The simulated response likelihoods quantify the likelihoods of the model-based unspeeded responses to end-of-test items. The ratio of the real to simulated response likelihoods provides a basis for quantifying the differences in

their distributions and the likelihood that particular responses are speeded. The smallest allowable ratio is then used to randomly eliminate examinee responses so that the relative distributions of response likelihoods are the same. This process is used to reduce the effects of speededness.

Table 1 and Figure 2a present hypothetical real and simulated response likelihood distributions for an item containing bias due to speededness. Table 1 and Figure 2b present the updated real response likelihood distribution after randomly eliminating data to reflect the simulated response likelihood distribution.

In Figure 2a, several bins have a larger frequency of real responses compared to simulated responses and several have a smaller frequency from this representation. The smallest bin ratio (real frequencies/simulated frequencies), .5, is used as the ratio that all real response likelihood bins must be reduced when randomly eliminating real responses from the data set. For each bin, examinees are randomly eliminated from the data set until the bin ratio is .5, which results in the comparable distributions observed in Figure 2b. The examinee responses to a given end-of-test item that were randomly eliminated are ultimately treated as missing and this updated item response data set calibrated.

**Real Data Study**

We used real data from a mathematics placement test (MPT) at a Midwestern university to illustrate and examine the proposed randomization procedure for reducing the effects of speededness. Test scores on the MPT are used to place university students into their first college mathematics course. The MPT consists of 36 five-option multiple-choice items. Two forms of the test, Form A and Form B, were examined in this study. Form A contained item responses for

10,157 examinees and Form B contained item responses for 2,496 examinees[1]. These two test forms were chosen because 11 items were common across Forms A and B but located at different points on the test. Four of the common items were located at the end of the test on Form A (items 31, 32, 33, and 34) but at earlier locations on Form B (items 4, 1, 3, and 2). This data structure allowed a real-data based evaluation of results using the randomization procedure in reducing bias for the end-of-test items on Form A.

An initial calibration of Form A and Form B data was completed using the computer program BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2002). The Rasch model was used to calibrate items, as this IRT model is used operationally for the MPT (see Appendix A for a description of the Rasch model). Form A item difficulty parameter estimates were scaled to Form B item difficulty parameter estimates using common item linking, test characteristic curve equating (Stocking and Lord, 1983), and the computer program EQUATE (Baker, 1993). Seven common test items in unspeeded locations on both test forms were used as the link. Table 2 lists the item parameter estimates for all of the common items across Forms A and B.

An examination of the percentage of omitted responses for Form A of the MPT showed that around 1% to 2% of item responses were omitted through much of the test and that these responses rose from 4% to 7% for items toward the end of the test, with a maximum of 7.2% of responses omitted for the last item (see the scatterplot in Figure 3). The percentage of omitted responses appeared to increase toward the end of the test, consistent with the presence of speededness effects. Because there was a larger proportion of omits at the end of the test than at the beginning and this proportion reflected a nontrivial number of omitted responses (a maximum of 7.2% or 730 examinees), it appeared that omits were useful to include as evidence

[1] Far fewer examinees took Form B of the MPT because it was a pilot form.

for speededness. Therefore, omits were included in the randomization procedure for the MPT data.

The randomization procedure was applied to Form A of the MPT, resulting in a data set for which aberrant responses were randomly excluded from the original item response data set. Each step of the procedure is reported below. The updated item response data set was then calibrated using the Rasch model and scaled to Form B to compare end-of-test item difficulties on Form A to the item difficulties of the same items in earlier locations on Form B.

Evaluation of the success of the procedure was considered in two ways. First, the differences in difficulty parameter estimates were examined across Form A and Form B before and after applying the randomization procedure. Second, the cumulative difference across items was compared by plotting the difference between test characteristic curves (TCCs) for the end-of-test items. TCCs display the expected sum scores across ability on the items that were in speeded locations on Form A and unspeeded locations on Form B. The difference between TCCs was computed as $TCC_A(q) - TCC_B(q)$, where $TCC_A(q)$ and $TCC_B(q)$ are the expected sum scores at ability $q$ for Form A and Form B, respectively. To summarize the difference in TCCs across Form A and Form B, an expected standardized difference index (ESDI) was calculated. The ESDI is the average squared difference between probabilities of correct response weighted by the distribution of ability. The equation for the ESDI is as follows:

$$ESDI = \frac{1}{\sum w(q)} \times \left( \sum\nolimits_q \left[ TCC_A(q) - TCC_B(q) \right]^2 \times w(q) \right) \qquad (2)$$

where

$w(q)$ is the weight based on the distribution of ability $q$,

$TCC_A(q)$ is the expected sum score for Form A end-of-test items at a given ability, and

$TCC_B(q)$ is the expected sum score for Form B end-of-test items at a given ability.

**Results**

The randomization procedure was applied to end-of-test items on Form A of the MPT. Each step of the procedure is described below.

*Step 1. Identify items that show evidence of speededness.* We identified items that showed evidence for speededness on the MPT in two ways. First, we examined the plot mentioned above displaying the percentage of omits for each item (see Figure 3) to observe the location on the test where omitting behavior appeared to become more frequent. Looking at Figure 3, the percentage of omits increases as item number increases (i.e., as we get closer to the end of the test). However, omits for the last four test items are particularly pronounced, with a percentage of omits around 6% to 7%. Also, between item 32 (five items from the end) and item 33 (four items from the end) there is a jump in percentage of omits from about 3.5% to almost 6%, which we might expect if examinees tend to become speeded at a similar point on the test. However, as mentioned above, considering only omits to identify items biased by speededness will likely underestimate the number of such items as examinees may also be rushed and/or guess on items.

The second method used to identify potential items biased by speededness was nonlinear factor analysis (via NOHARM; Fraser, 1988). The factor loadings from an exploratory two-factor solution are listed in Table 3. All items load highly on the first factor, consistent with a primary factor underlying all test items. The last six items (and only the last six items) resulted in large loadings for a second factor, consistent with a speededness factor. Thus, the last six items were identified as the end-of-test items for which the effects of speededness would be reduced.

*Step 2. Fit a nominal response model (NRM).* A NRM was fit to the data set (using MULTILOG; Thissen, Chen, & Bock, 2003), fixing the omit category to zero for the end-of-test

items identified in Step 1. The resultant NRM parameter estimates are listed in Appendix B. An estimated ORF for the omit category was added to the end-of-test items later using the procedure described earlier.

*Step 3. Obtain an estimate of examinee ability uncontaminated by speededness.* Examinee ability estimates were based on unspeeded items only, to reduce potential bias due to speededness. For the MPT data, the last six items showed evidence of speededness, so these items were not used for ability estimation. These estimates of examinee ability were later used to estimate real and simulated response likelihoods for each examinee to each item (described below).

*Steps 4 and 5. Construct an option response function (ORF) for the omit category, consistent with unspeeded responding. Apply the constructed unspeeded omit ORF to the OCCs of end-of-test items.* To approximate an unspeeded ORF for the omitted responses, the average ORF for the omit category across all unspeeded items was constructed. This ORF was constructed using the omitted response NRM parameter estimates for each unspeeded item and a fixed set of points on the ability scale. The first 30 items on the test were used as a basis for this unspeeded omit ORF. Defining the unspeeded omit ORF differently may lead to different omit ORFs and ultimately to different results. For example, the ORF could be based on the average of the omit ORFs for the first several test items. However, in the MPT data, such an ORF would differ little from one based on the first 30 items because the probability of an omitted response is very small across all unspeeded items. The OCC corresponding to this ORF is displayed in Figure 4. This average OCC was then applied to the NRM OCCs for each end-of-test item estimated in step 2. These OCCs are used as the basis for calculating real and simulated response likelihoods and are listed in Figures 5a through 5f.

*Steps 6, 7, 8, and 9. For end-of-test items, compute the likelihood of each observed response. Simulate a model-based set of examinee responses and compute the likelihood of each model-based response. Segment the response likelihoods into bins. Compute ratios of real to simulated response likelihoods for each bin*. Recall that the observed response likelihoods are the NRM likelihoods of the response category chosen by each examinee and that the simulated response likelihoods are the NRM likelihoods of the response category based on simulated responses. Each of these response likelihoods was calculated for the real and simulated data and distributions of these response likelihoods are displayed in Figures 6a through 6f. In addition, ratios of real to simulated responses for each bin of each end-of-test item are displayed in Table 4. Of note, bin number 1, which reflects small response likelihoods, had a ratio greater than one for each of the six end-of-test items. The bin ratio for bin number 1 in each of the end-of-test items was 1.02, 1.23, 1.31, 1.54, 1.93, and 1.46, respectively. This means that a larger number of examinees responded to end-of-test items in a way that resulted in smaller response likelihoods than would be expected under the NRM.

*Step 10. Set a minimum ratio for the real to simulated response likelihoods (to ensure adequate bin sizes) and randomly eliminate item responses in proportion to this ratio.* Response likelihood distributions based on real and simulated data after randomly eliminating examinee responses are displayed in Figures 7a through 7f. Notice that for each item, the real response likelihood distributions are all proportional to the real response likelihood distributions.

Ultimately, we are interested in using the randomization procedure to reduce bias in IRT item parameter estimates of end-of-test items for the IRT model used to calibrate and/or score the test. Table 5 lists the common updated Rasch model difficulty parameters for Form A after scaling them to Form B MPT parameter estimates. For the four common items at the end of the

test on Form A and at the beginning of the test on Form B, the item difficulties for three out of four of the items were closer to Form B estimates after applying the randomization procedure. All 36 item difficulty parameter estimates before and after applying the randomization procedure are listed in Appendix C.

We also examined the reduction in bias for the four items together. Differences between test characteristic curves for Form A and Form B on the four common items *before* applying the randomization procedure are displayed in Figure 8a. The ESDI between Form A and Form B on these four items is 0.018. Differences between test characteristic curves for Form A and Form B on the four common items *after* applying the randomization procedure are displayed in Figure 8b. The ESDI between Form A and Form B on these four items is 0.005. Therefore, the difference quantified by the ESDI was reduced by over 70%.

**Discussion**

Speededness is a potential source of bias on item and test characteristics. Methods for identifying and reducing speededness effects are important for addressing this potential source of bias and to ensure a better evaluation of how the items will perform when administered at different locations on new forms. This study presented one practical IRT-based method for reducing speededness effects that considers the likelihood of response under the NRM as a basis for probabilistically eliminating aberrant responses. The procedure can be applied to data sets having large numbers of examinees and items, typical of large testing programs.

The results of applying the procedure to the MPT showed that item parameter estimates tended to be closer to those item parameter estimates observed when the items were administered at earlier locations on a different test form. Also, the difference in expected sum scores across end-of-test items was reduced when applying the randomization procedure.

For researchers already familiar with methods of person fit, this study illustrates an application of a method that falls within the same general framework but that considers (a) unordered item response choices via the NRM and (b) a specific type of response aberrance (speededness). It is important to note that the method is not restricted to studying speededness. If some number of identifiable items is suspected of containing bias for a known reason, the randomization procedure could be applied in a similar way to reduce the bias. For example, if a subset of examinees carelessly responded to long items (i.e., items with many words) by ignoring the questions and picking the first response option that appeared reasonable, the randomization procedure may be useful to apply.

A potential limitation of the randomization procedure is that the end-of-test NRM item parameter estimates for all categories except omits, are biased by speededness. However, these parameter estimates are the basis for constructing response likelihood distributions and randomly eliminating responses. As such, the procedure carries bias due to speededness through the process of reducing the effects of speededness, which may affect its ability to minimize bias due to speededness. A possible solution is to iteratively apply the procedure by using the updated item parameter estimates after applying the randomization procedure once as the initial item parameter estimates in a subsequent application of the procedure.

A limitation of this study is the generalizability of the randomization procedure to other data sets, as we have only illustrated its utility using one real data set. Simulation studies will be useful for examining how well the procedure works under a range of conditions when the true underlying item and test characteristics are known. We are currently undertaking one such study.

## References

Baker, F. B. (1993). Equate 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement, 17,* 20.

Bejar, I. I. (1985). *Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language.* ETS Research Report RR-85-11.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

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

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

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists.* Lawrence Erlbaum Associates: Mahwah, New Jersey.

Fraser, C. (1988). *NOHARM*. [Computer Software]. New South Wales, Australia: Author.

Fox, J.–P. (2005). Randomized item response theory models. *Journal of Educational and Behavioral Statistics, 30,* 189-212.

Kingston, N. M. & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8,* 147-154.

Lord, F. M. (1956). A study of speed factors in tests and academic grades. *Psychometrika, 21,* 31-50.

Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Addison-Wesley: Reading, MA.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Psychological Measurement, 9,* 3-8.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8,* 72-87.

Meijer, R. R. & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25,* 107-135.

Mroch, A. A., Bolt, D. M., & Wollack, J. A. (2005). *A new multiclass mixture Rasch model for test speededness.* Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Quebec.

Myers, C. T. (1952). The factorial composition and validity of differently speeded tests. *Psychometrika, 17,* 347-352.

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

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Tatsuoka, K. (1996). Use of generalized person-fit indexes, zetas for statistical pattern classification. *Applied Measurement in Education, 9,* 65-75.

Thissen, D., Chen, W.-H., & Bock, D. (2003). *MULTILOG (version 7)* [Computer software]. Lincolnwood, IL: Scientific Software International.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60,* 63-69.

Yamamoto, K. (1987). *A model that combines IRT and latent class models.* Unpublished doctoral dissertation. University of Illinois, Champaign – Urbana.

Yamamoto, K. & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost and R. Langeheine (Eds.): *Applications of Latent Trait and Latent Class Models in the Social Sciences* (pp. 89 – 98). New York: Waxman.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2002). *BILOG-MG (version 3)* [Computer software]. Lincolnwood, IL: Scientific Software International.

Table 1
Example real and simulated response likelihood bin counts and ratios for a hypothetical item

| Bin | Observed Count | Model-based Count | Ratio (Obs./Model) | Observed Count after random elimination | Ratio after random elimination |
|---|---|---|---|---|---|
| .1 - .2 | 500 | 500 | 1 | 250 | 0.5 |
| .2 - .3 | 1000 | 500 | 2 | 250 | 0.5 |
| .3 - .4 | 1000 | 500 | 2 | 250 | 0.5 |
| .4 - .5 | 1500 | 1000 | 1.5 | 500 | 0.5 |
| .5 - .6 | 1500 | 1500 | 1 | 750 | 0.5 |
| .6 - .7 | 1500 | 2000 | 0.75 | 1000 | 0.5 |
| .7 - .8 | 1500 | 3000 | 0.5* | 1500 | 0.5 |
| .8 - .9 | 1000 | 500 | 2 | 250 | 0.5 |
| .9 - 1 | 500 | 500 | 1 | 250 | 0.5 |

*Smallest ratio value; used as the ratio for random elimination of data in observed counts.

Table 2
Item parameter estimates for common items across Form A and Form B of the mathematics placement test

| Item Number Math Form B | Form B Difficulty Parameter Estimate | Item Number Math Form A | Form A Difficulty Parameter Estimate | Difference in Difficulty Parameter Estimates |
|---|---|---|---|---|
| 20 | 0.36 | 11 | 0.56 | 0.20 |
| 12 | 0.66 | 16 | 0.49 | -0.17 |
| 5 | -0.42 | 22 | -0.43 | -0.01 |
| 15 | 0.19 | 23 | 0.23 | 0.04 |
| 19 | -0.28 | 25 | -0.25 | 0.03 |
| 28 | 0.37 | 27 | 0.27 | -0.10 |
| 22 | -0.09 | 30 | -0.07 | 0.02 |
| 4 | -0.79 | 31 | -0.64 | 0.15 |
| 1 | -1.51 | 32 | -1.09 | 0.42 |
| 3 | -0.31 | 33 | -0.26 | 0.05 |
| 2 | -0.04 | 34 | 0.07 | 0.11 |

Note: The items shaded in grey were not used to link Form A to Form B. These items can be used to examine the bias in end-of-test items for Form A. The item in the last row is located at the end of the test on both forms, making bias examination difficult.

Table 3
Two-factor exploratory factor analysis solution for MPT Form A

| Item | Factor 1 | Factor 2 |
|---|---|---|
| 1 | 0.471 | 0.000 |
| 2 | 0.517 | -0.009 |
| 3 | 0.463 | 0.028 |
| 4 | 0.418 | 0.013 |
| 5 | 0.522 | -0.019 |
| 6 | 0.545 | -0.049 |
| 7 | 0.636 | -0.067 |
| 8 | 0.580 | -0.138 |
| 9 | 0.563 | -0.037 |
| 10 | 0.547 | -0.006 |
| 11 | 0.477 | -0.086 |
| 12 | 0.414 | -0.058 |
| 13 | 0.337 | 0.080 |
| 14 | 0.443 | 0.159 |
| 15 | 0.329 | 0.082 |
| 16 | 0.448 | -0.028 |
| 17 | 0.202 | -0.050 |
| 18 | 0.567 | -0.032 |
| 19 | 0.573 | 0.102 |
| 20 | 0.566 | -0.023 |
| 21 | 0.512 | 0.070 |
| 22 | 0.508 | 0.052 |
| 23 | 0.585 | 0.003 |
| 24 | 0.546 | -0.046 |
| 25 | 0.468 | 0.165 |
| 26 | 0.381 | 0.203 |
| 27 | 0.570 | 0.089 |
| 28 | 0.643 | 0.246 |
| 29 | 0.431 | 0.091 |
| 30 | 0.388 | 0.150 |
| 31 | 0.345 | 0.414 |
| 32 | 0.501 | 0.567 |
| 33 | 0.473 | 0.371 |
| 34 | 0.575 | 0.255 |
| 35 | 0.528 | 0.407 |
| 36 | 0.589 | 0.249 |

Table 4
Observed to expected ratios in each response likelihood bin for end-of-test items on the mathematics placement test.

| Bin* | Item Number | | | | | |
|---|---|---|---|---|---|---|
| | **31** | **32** | **33** | **34** | **35** | **36** |
| 1 | 1.02 | 1.23 | 1.31 | 1.54 | 1.93 | 1.46 |
| 2 | 0.72** | 1.00 | 0.93 | 0.92 | 1.00 | 0.99 |
| 3 | 1.12 | 0.88** | 0.86** | 1.02 | 0.90 | 0.93 |
| 4 | 1.04 | 0.93 | 1.01 | 0.89 | 0.90 | 0.91 |
| 5 | 0.98 | 0.96 | 1.01 | 0.89 | 0.92 | 0.85 |
| 6 | 1.01 | 0.97 | 0.93 | 0.85** | 0.86** | 0.85** |
| 7 | -- | -- | 0.90 | 0.97 | 0.90 | 0.97 |
| 8 | -- | -- | 0.92 | 0.93 | 0.91 | 0.94 |
| 9 | -- | -- | 0.99 | 1.01 | 0.95 | 1.01 |
| 10 | -- | -- | 1.04 | 0.98 | 0.99 | 0.99 |

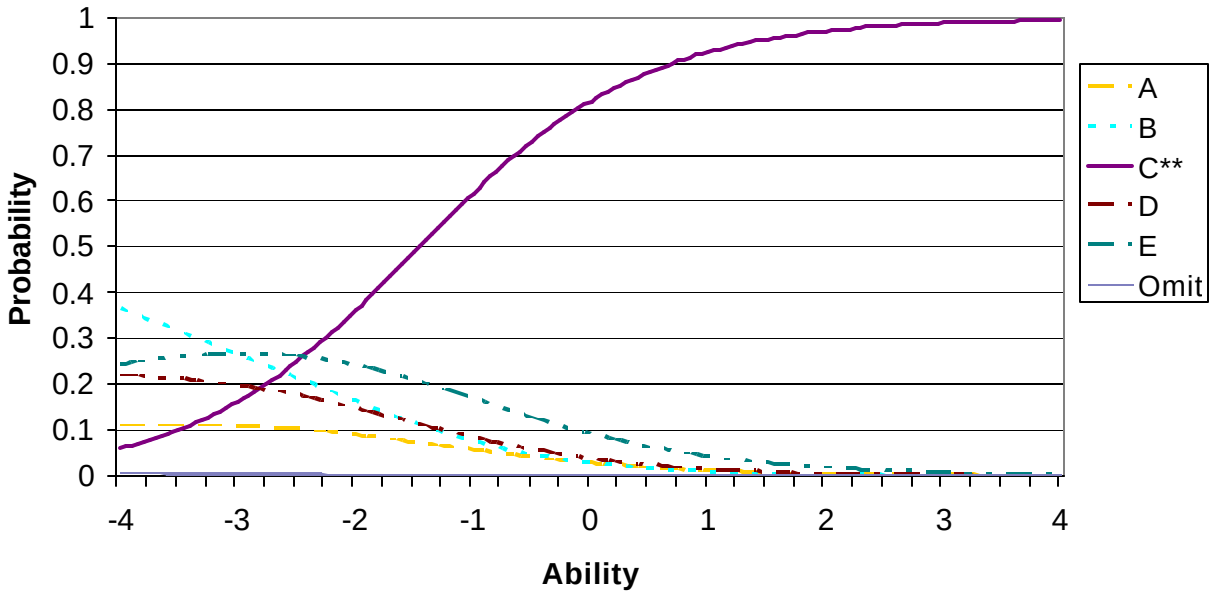
*Note that the number of bins varied by item to ensure adequate bin sizes. The range of response likelihoods reflected by each bin corresponds to one divided by the number of bins (e.g., for item 31, the bin width is 1/6 or .167.
**Smallest ratio value; used as the ratio for random elimination of data in observed counts.

Table 5
Item parameter estimates for common items across Form A and Form B of the Mathematics Placement Test before and after applying the randomization procedure to Form A end-of-test items.

| Item Number Form B | Form B Difficulty Parameter Estimate | Item Number Form A | Form A Difficulty Parameter Estimate | Form A Difficulty Parameter Estimate After Applying Randomization Procedure |
|---|---|---|---|---|
| 20 | 0.36 | 11 | 0.56 | 0.56 |
| 12 | 0.66 | 16 | 0.49 | 0.49 |
| 5 | -0.42 | 22 | -0.43 | -0.43 |
| 15 | 0.19 | 23 | 0.23 | 0.24 |
| 19 | -0.28 | 25 | -0.25 | -0.25 |
| 28 | 0.37 | 27 | 0.27 | 0.27 |
| 22 | -0.09 | 30 | -0.07 | -0.07 |
| 4 | -0.79 | 31 | -0.64 | -0.64 |
| 1 | -1.51 | 32 | -1.09 | -1.23 |
| 3 | -0.31 | 33 | -0.26 | -0.35 |
| 2 | -0.04 | 34 | 0.07 | -0.02 |

Note: The items shaded in grey were not used to link Form A to Form B.

Figure 1
Example Option Characteristic Curves for a Nominal Response Model with Six Categories




| NRM Parameter | Item Response Option | | | | | |
|---|---|---|---|---|---|---|
| | A | B | C** | D | E | Omit |
| a | 0.0 | -0.3 | 1.0 | -0.1 | 0.1 | -0.7 |
| c | 0.9 | 0.9 | 4.3 | 1.2 | 2.1 | -5.0 |

**Correct response.

Figure 2a
Plot of example real and simulated response likelihood bin counts for a hypothetical item.

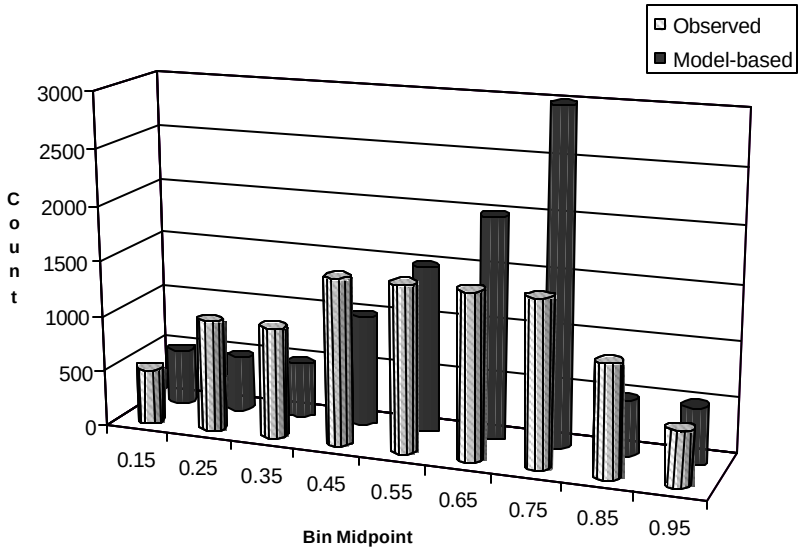



Figure 2b
Plot of example real and simulated response likelihood bin counts after randomly eliminating data from observed bins for a hypothetical item.

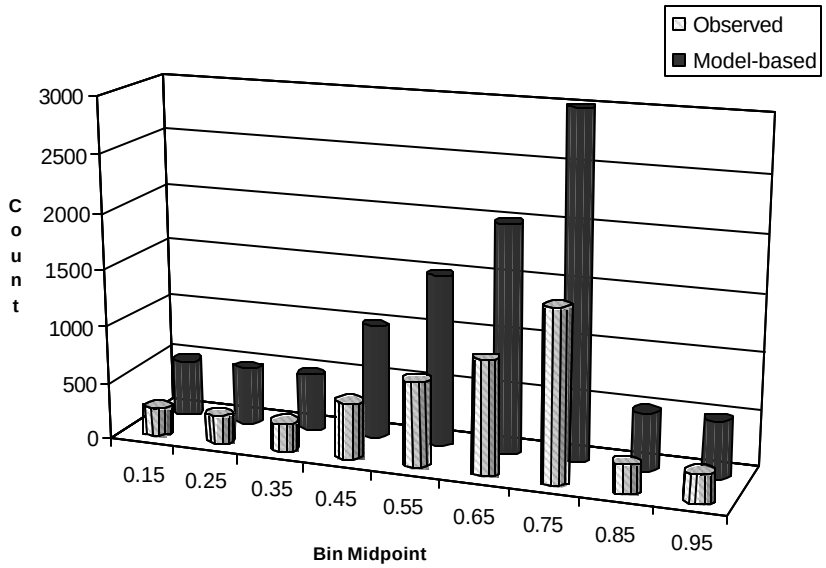


Figure 3
Scatterplot of percentage of omits by item number for Form A of the mathematics placement test

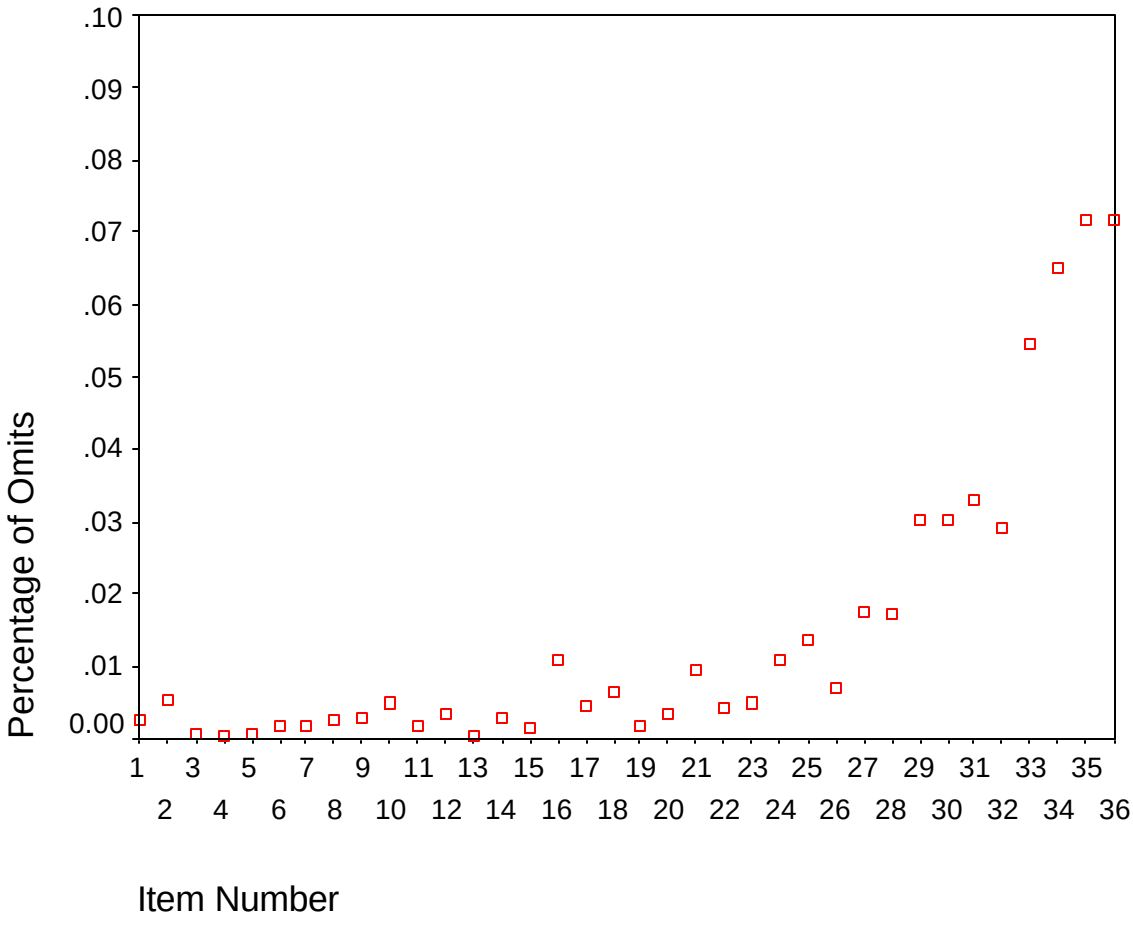


Figure 4
OCC for constructed unspeeded omit ORF for the mathematics placement test.

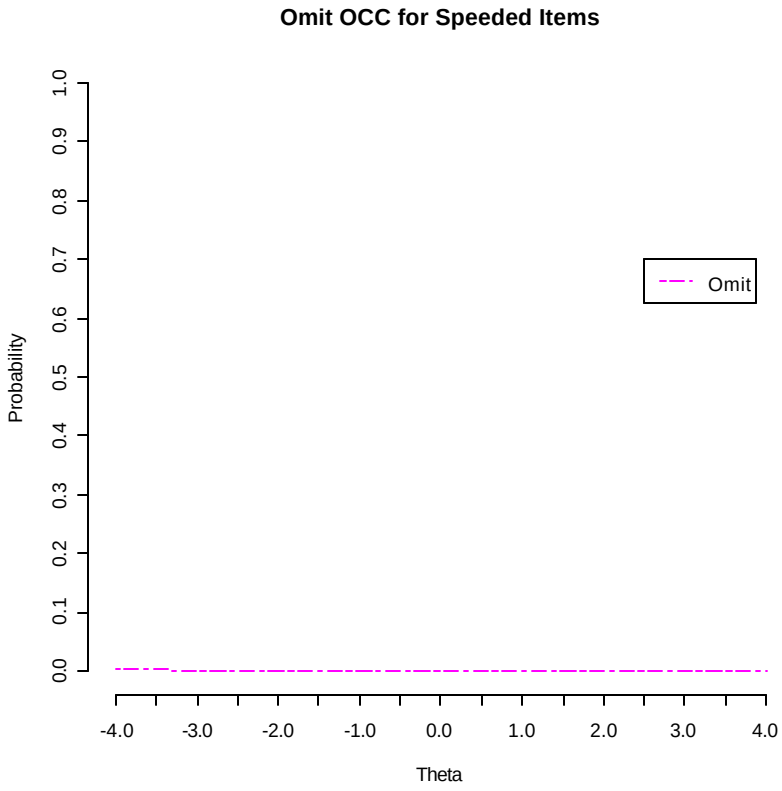


Figure 5a
OCCs for item 31 of the mathematics placement test from step 5 of the randomization procedure.

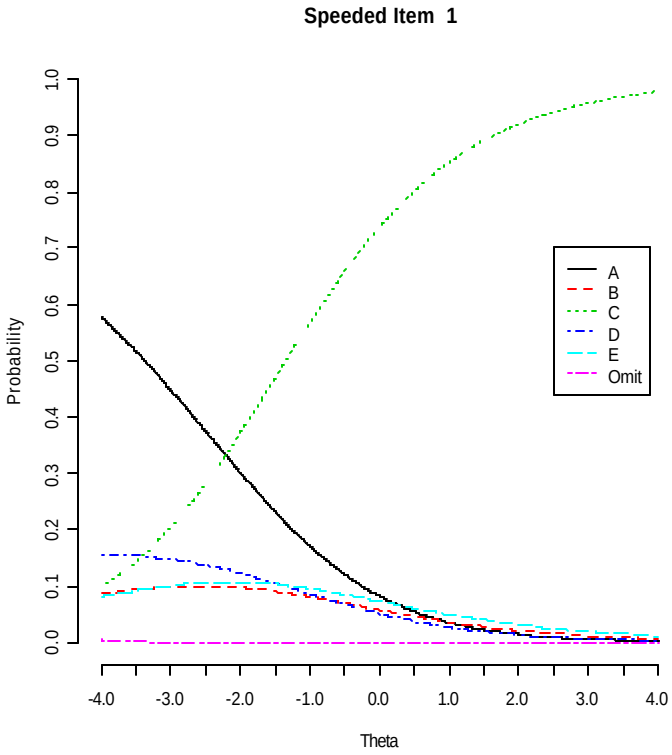



Figure 5b
OCCs for item 32 of the mathematics placement test from step 5 of the randomization procedure.

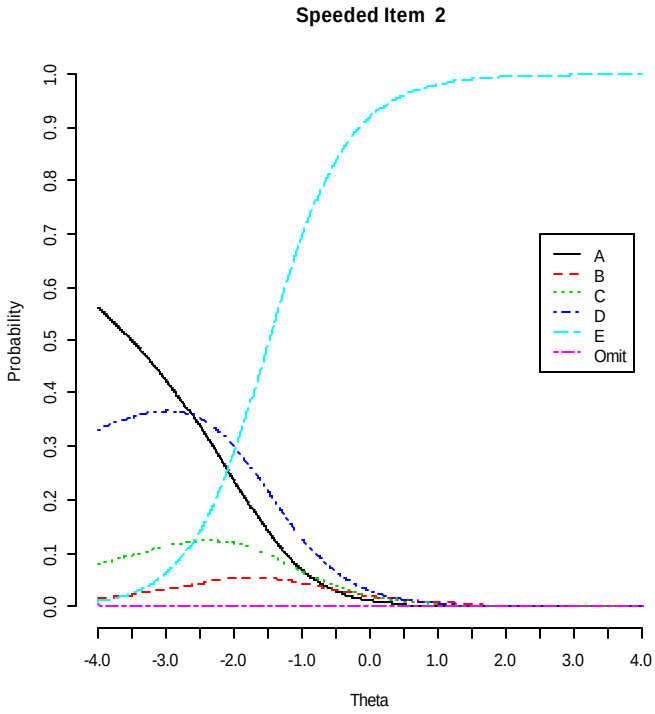


Figure 5c
OCCs for item 33 of the mathematics placement test from step 5 of the randomization procedure.

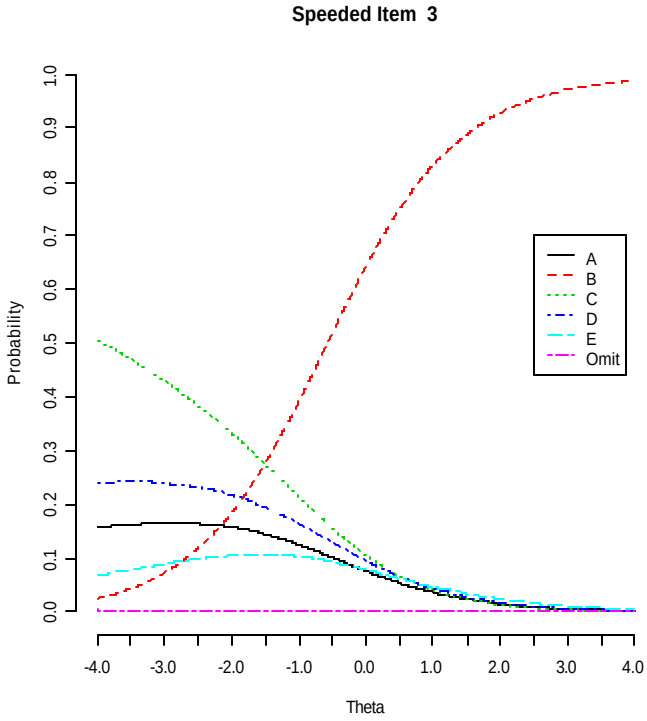



Figure 5d
OCCs for item 34 of the mathematics placement test from step 5 of the randomization procedure.

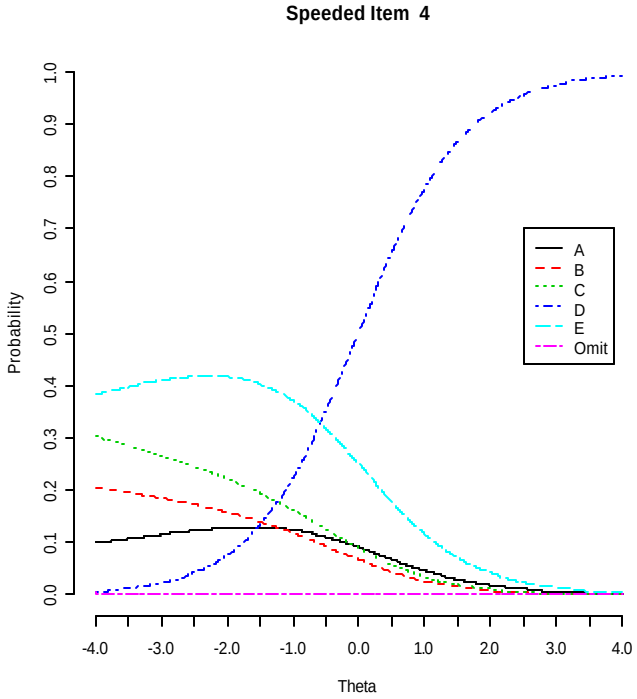


Figure 5e
OCCs for item 35 of the mathematics placement test from step 5 of the randomization procedure.

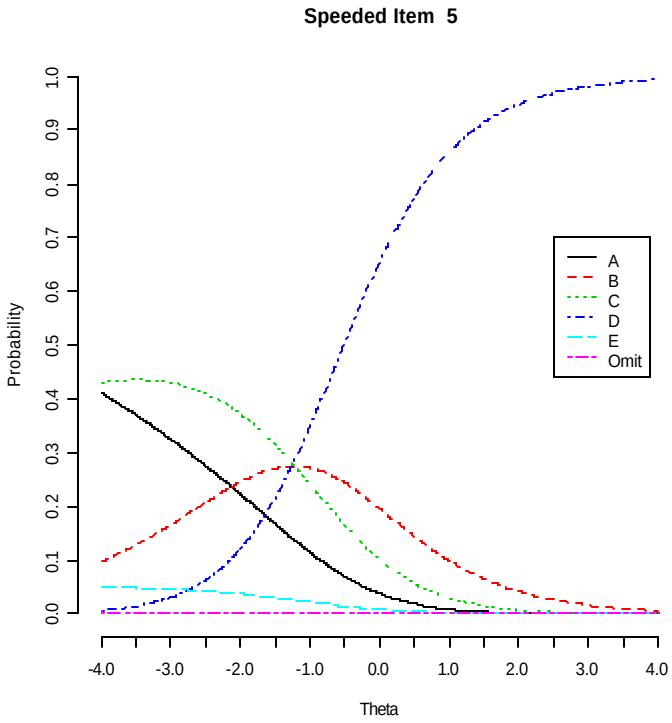



Figure 5f
OCCs for item 36 of the mathematics placement test from step 5 of the randomization procedure.

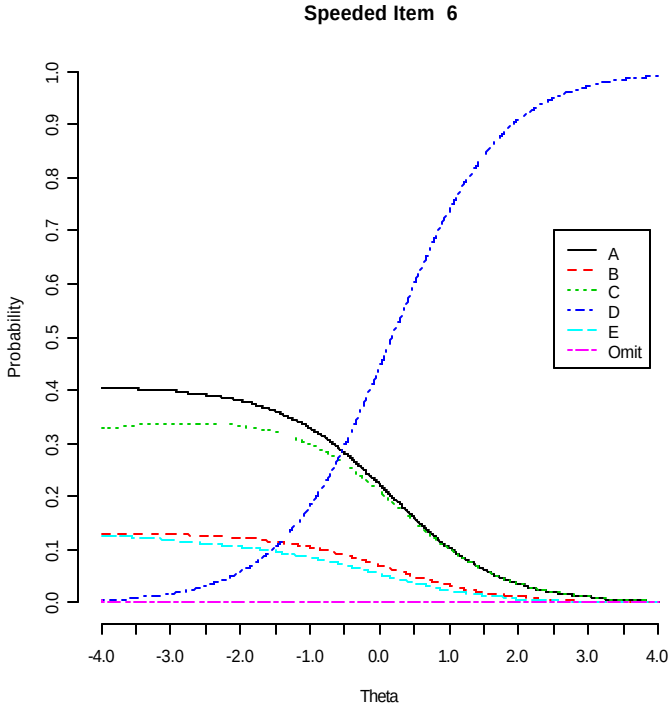


Figure 6a
Real and simulated response likelihood bins for item number 31 of the mathematics placement test.

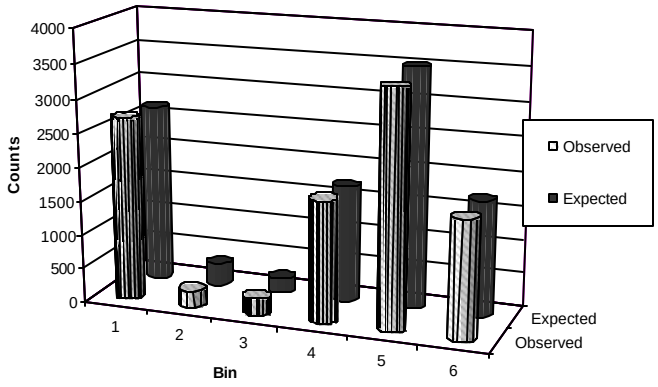



Figure 6b
Real and simulated response likelihood bins for item number 32 of the mathematics placement test.

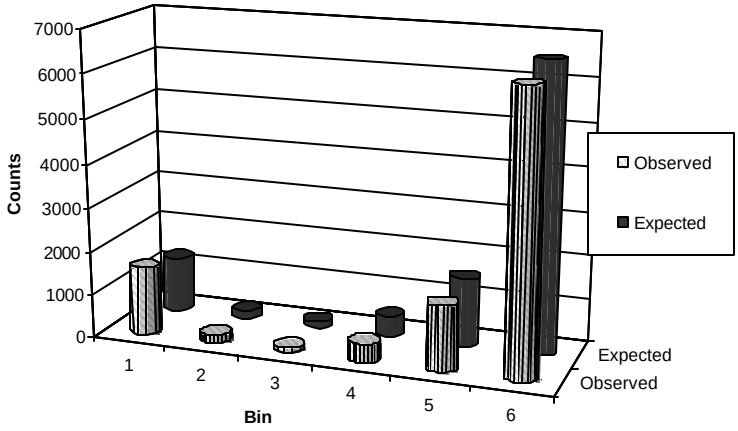



Figure 6c
Real and simulated response likelihood bins for item number 33 of the mathematics placement test.

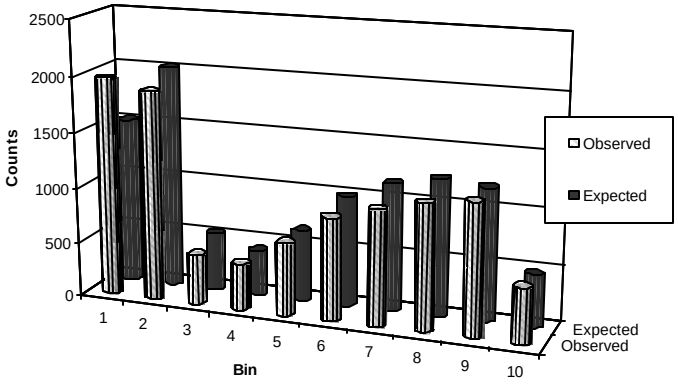


Figure 6d
Real and simulated response likelihood bins for item number 34 of the mathematics placement test.

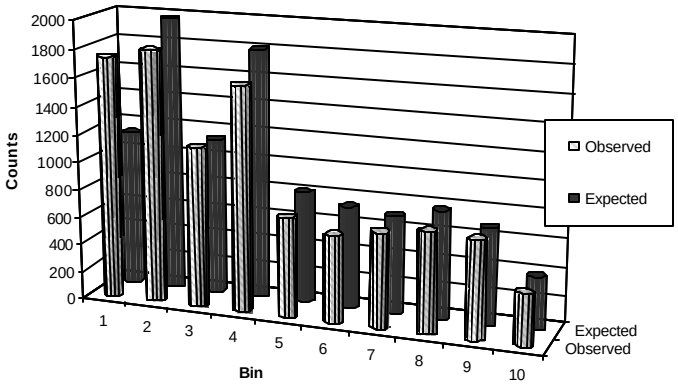



Figure 6e
Real and simulated response likelihood bins for item number 35 of the mathematics placement test.

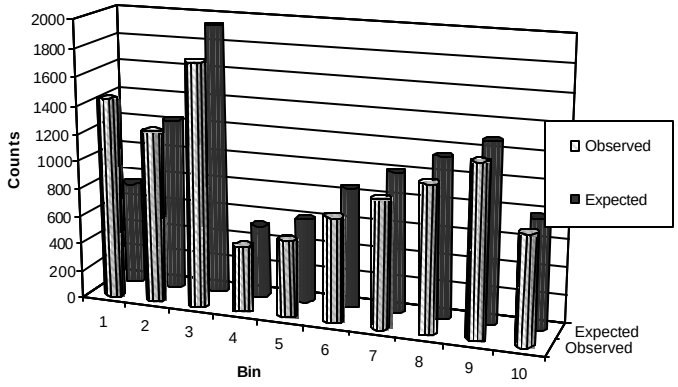



Figure 6f
Real and simulated response likelihood bins for item number 36 of the mathematics placement test.

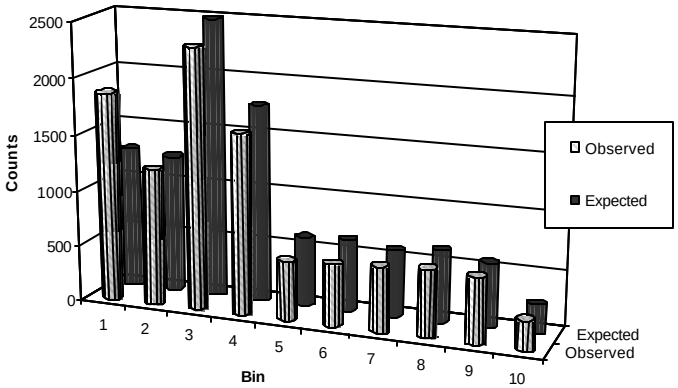


Figure 7a
Real and simulated response likelihood bins for item number 31 of the mathematics placement test after randomly eliminating aberrant responses.

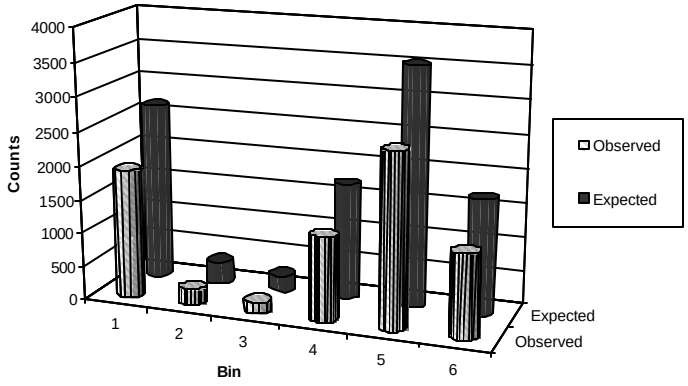



Figure 7b
Real and simulated response likelihood bins for item number 32 of the mathematics placement test after randomly eliminating aberrant responses.

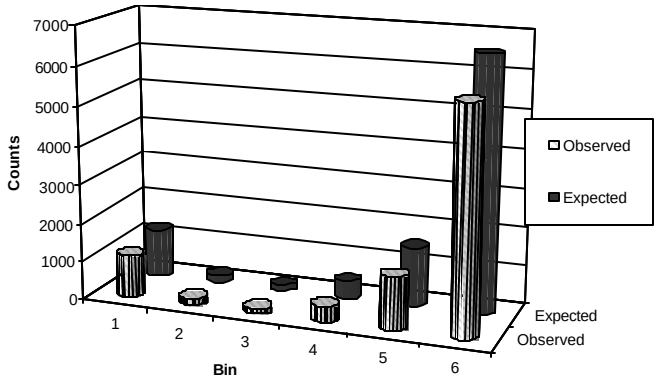



Figure 7c
Real and simulated response likelihood bins for item number 33 of the mathematics placement test after randomly eliminating aberrant responses.

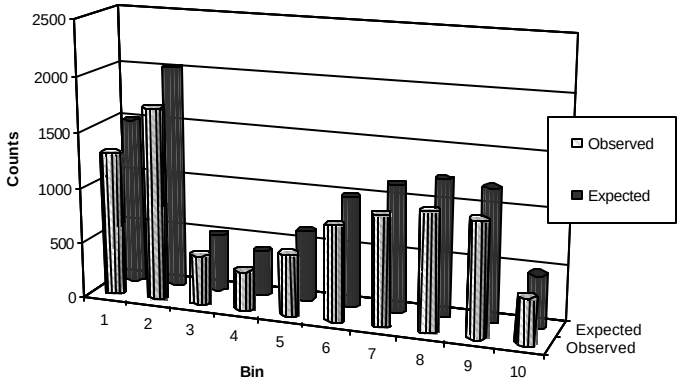


Figure 7d
Real and simulated response likelihood bins for item number 34 of the mathematics placement test after randomly eliminating aberrant responses.

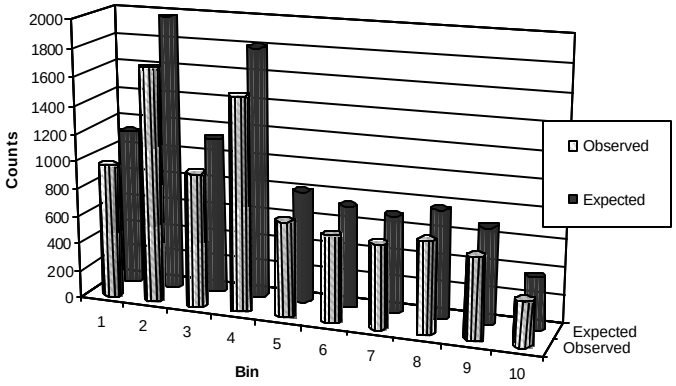



Figure 7e
Real and simulated response likelihood bins for item number 35 of the mathematics placement test after randomly eliminating aberrant responses.

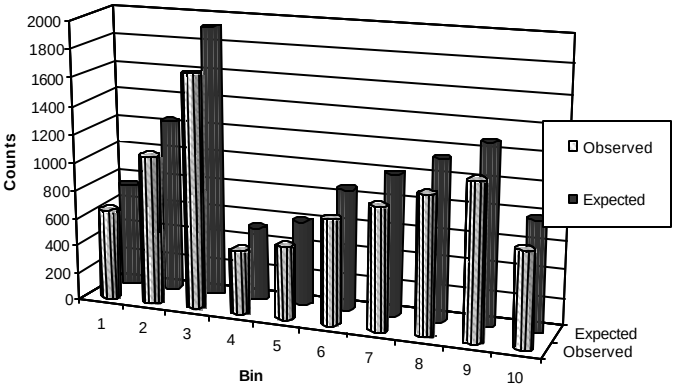



Figure 7f
Real and simulated response likelihood bins for item number 36 of the mathematics placement test after randomly eliminating aberrant responses.

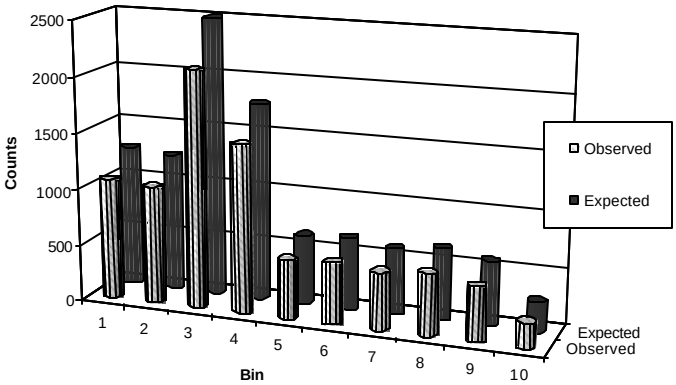


Figure 8a
Difference in TCCs for four common items in end-of-test locations on Form A and beginning-of-test locations on Form B before applying the randomization approach (ESDI: .018)

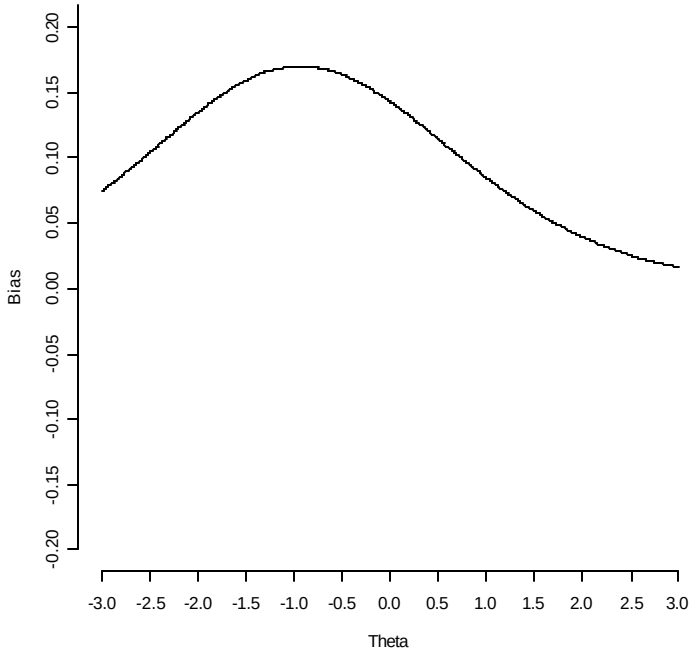
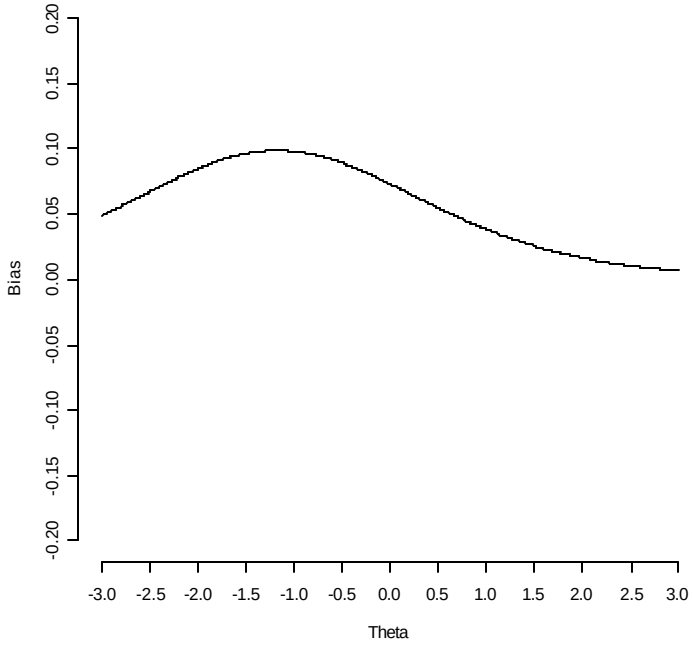



Figure 8b
Difference in TCCs for four common items in end-of-test locations on Form A and beginning-of-test locations on Form B after applying the randomization approach (ESDI: .005)



## Appendix A
## Description of the Rasch Model

The Rasch model (Rasch, 1960) is a dichotomous IRT model that identifies the probability of examinee *j* answering item *i* correctly as $P(u_{ij}=1|q_j, b_i)=\frac{e^{(q_j-b_i)}}{1+e^{(q_j-b_i)}}$, where $u_{ij}$ is the response to the item (0 = incorrect, 1 = correct), $b_i$, is the item difficulty parameter, and $q_j$ is the examinee ability parameter.

For each item, the probability that an examinees at various abilities will respond correctly to an item can be plotted (this plot is commonly referred to as an *item characteristic curve* or *ICC*). A Rasch model ICC for one item (with difficulty parameter equal to zero) is presented in the figure below. This ICC can be used to identify the probability that a given examinee will answer an item correctly. For example, an examinee with ability 0 has a .5 probability of correctly responding to the item plotted below.

**Appendix B**
**NRM parameter estimates for Form A of the mathematics placement test**

| Item | NRM a Parameter Estimate | | | | | |
|---|---|---|---|---|---|---|
| | a1 | a2 | a3 | a4 | a5 | a6 |
| 1 | -0.17 | -0.53 | 0.81 | -0.31 | -0.07 | 0.25 |
| 2 | -0.40 | -0.60 | -0.31 | 0.19 | 1.05 | 0.01 |
| 3 | -0.38 | -0.41 | 0.85 | 0.01 | 0.01 | -0.13 |
| 4 | -0.64 | -0.22 | 0.69 | -0.40 | 0.38 | 0.11 |
| 5 | -0.40 | 1.03 | -0.08 | 0.18 | -0.39 | -0.39 |
| 6 | -0.66 | -0.16 | -0.31 | -0.03 | 0.83 | 0.25 |
| 7 | -0.19 | -0.56 | 1.06 | -0.14 | -0.24 | 0.04 |
| 8 | 0.32 | -0.53 | -0.53 | 1.02 | -0.23 | -0.02 |
| 9 | 0.89 | -0.08 | -0.16 | -0.16 | -0.58 | 0.20 |
| 10 | -0.51 | -0.32 | -0.18 | 0.87 | 0.30 | -0.21 |
| 11 | -0.59 | 0.71 | -0.19 | -0.46 | 0.17 | 0.30 |
| 12 | 0.63 | -0.33 | -0.61 | -0.20 | 0.11 | 0.46 |
| 13 | 0.11 | -0.84 | 0.58 | -0.09 | -0.37 | 0.63 |
| 14 | -0.61 | -0.55 | -0.05 | 0.72 | 0.46 | -0.03 |
| 15 | 0.55 | -0.26 | -0.05 | -0.07 | -0.30 | 0.19 |
| 16 | 0.22 | 0.70 | -0.06 | -0.25 | -0.22 | -0.36 |
| 17 | -0.45 | 0.23 | -0.18 | 0.22 | -0.18 | 0.31 |
| 18 | 1.15 | 0.21 | -0.50 | -0.05 | 0.02 | -0.70 |
| 19 | -0.25 | -0.21 | -0.51 | 1.11 | -0.12 | -0.04 |
| 20 | -0.36 | -0.07 | -0.89 | 0.78 | -0.11 | 0.60 |
| 21 | 1.15 | -0.10 | 0.29 | -0.03 | 0.04 | -1.21 |
| 22 | 0.46 | -0.25 | 0.21 | 1.28 | -0.38 | -1.27 |
| 23 | 0.11 | 1.29 | 0.21 | -0.14 | 0.30 | -1.76 |
| 24 | -0.28 | -0.69 | -0.10 | 0.86 | 0.27 | -0.10 |
| 25 | -0.31 | 0.65 | -0.21 | -0.30 | -0.37 | 0.51 |
| 26 | 0.80 | 0.02 | -0.10 | -0.59 | 0.03 | -0.06 |
| 27 | 0.96 | -0.19 | -0.47 | 0.04 | -0.38 | 0.15 |
| 28 | 1.42 | -0.29 | -0.57 | -0.59 | -0.17 | 0.37 |
| 29 | 0.32 | 0.79 | -0.29 | -0.01 | -0.23 | -0.54 |
| 30 | -0.69 | -0.11 | 0.02 | 0.60 | -0.04 | 0.14 |
| 31 | -0.45 | -0.07 | 0.54 | -0.24 | 0.01 | * |
| 32 | -0.89 | 0.17 | -0.25 | -0.51 | 1.23 | * |
| 33 | -0.20 | 0.78 | -0.41 | -0.25 | 0.01 | * |
| 34 | -0.11 | -0.37 | -0.39 | 1.01 | -0.19 | * |
| 35 | -0.59 | 0.17 | -0.36 | 1.13 | -0.44 | * |
| 36 | -0.23 | -0.24 | -0.20 | 1.05 | -0.30 | * |

*Parameters for end-of-test item omitted responses were fixed.

**Appendix B (continued)**
**NRM parameter estimates for Form A of the mathematics placement test**

| Item | NRM c Parameter Estimate | | | | | |
|---|---|---|---|---|---|---|
| | c1 | c2 | c3 | c4 | c5 | c6 |
| 1 | 1.09 | 1.03 | 4.41 | 1.30 | 2.24 | -9.95 |
| 2 | -0.81 | -0.32 | 0.11 | 1.32 | 2.94 | -3.35 |
| 3 | 0.88 | 0.26 | 3.63 | 0.91 | 0.03 | -5.60 |
| 4 | 1.24 | 2.59 | 4.32 | 0.07 | 1.81 | -9.87 |
| 5 | 1.97 | 3.74 | 0.81 | 3.36 | 0.40 | -10.05 |
| 6 | 0.70 | 0.95 | 0.17 | 1.45 | 2.25 | -5.44 |
| 7 | 3.16 | 1.69 | 3.47 | 1.65 | 0.68 | -10.28 |
| 8 | 2.73 | 2.21 | 1.14 | 2.59 | 1.42 | -9.78 |
| 9 | 3.00 | 2.37 | 2.09 | 2.73 | 0.92 | -10.75 |
| 10 | 1.08 | 2.18 | 2.41 | 3.03 | 1.27 | -9.84 |
| 11 | 0.70 | 1.29 | 0.84 | -0.67 | 1.34 | -3.41 |
| 12 | 2.21 | 1.30 | -1.01 | 0.09 | 1.76 | -4.10 |
| 13 | 0.94 | -0.73 | 3.54 | 0.78 | -0.02 | -4.40 |
| 14 | 0.86 | 1.91 | 2.27 | 3.83 | 1.11 | -9.88 |
| 15 | 5.19 | 1.23 | 1.76 | 4.25 | 1.02 | -12.83 |
| 16 | 0.77 | 1.59 | 1.27 | 0.85 | 0.80 | -5.18 |
| 17 | -0.58 | 1.88 | 1.45 | 0.89 | 0.57 | -4.29 |
| 18 | 1.86 | 2.17 | 0.59 | 0.15 | 0.91 | -5.46 |
| 19 | 1.52 | 2.67 | 0.78 | 4.06 | 0.61 | -9.45 |
| 20 | 0.62 | 0.03 | 0.98 | 2.01 | 1.92 | -5.49 |
| 21 | 3.03 | 1.43 | 1.30 | 0.12 | 0.75 | -6.27 |
| 22 | 1.85 | 0.18 | 1.64 | 3.32 | -0.46 | -6.31 |
| 23 | 1.25 | 2.54 | 2.44 | 0.92 | 0.53 | -7.52 |
| 24 | 0.16 | 0.85 | 2.22 | 1.62 | 0.64 | -5.47 |
| 25 | -0.44 | 2.55 | 1.44 | 0.46 | 0.67 | -4.72 |
| 26 | 3.45 | 0.12 | 0.57 | -1.57 | 1.37 | -3.54 |
| 27 | 2.90 | 1.35 | 1.59 | 2.58 | 1.97 | -10.05 |
| 28 | 4.20 | 1.61 | 1.30 | 1.02 | 1.91 | -9.55 |
| 29 | 1.46 | 1.90 | 1.35 | 1.00 | 0.26 | -5.79 |
| 30 | -0.32 | 0.80 | 0.51 | 2.04 | 0.85 | -3.92 |
| 31 | 0.69 | 0.33 | 2.89 | 0.21 | 0.57 | * |
| 32 | 0.58 | 1.20 | 1.17 | 1.59 | 5.03 | * |
| 33 | 1.58 | 3.71 | 1.91 | 1.81 | 1.61 | * |
| 34 | 1.58 | 1.27 | 1.56 | 3.29 | 2.59 | * |
| 35 | 1.16 | 2.80 | 2.14 | 4.00 | -0.31 | * |
| 36 | 2.42 | 1.27 | 2.36 | 3.11 | 1.01 | * |

*Parameters for end-of-test omitted responses were fixed.

## Appendix C
## Rasch Model Item Difficulty Parameter Estimates for Form A of the Mathematics Placement Test Before and After Applying the Randomization Procedure

| | Before Applying the Randomization Procedure | | After Applying the Randomization Procedure | |
|---|---|---|---|---|
| Item | Rasch Difficulty | Standard Error | Rasch Difficulty | Standard Error |
| 1 | -0.86 | 0.03 | -0.87 | 0.03 |
| 2 | -0.64 | 0.02 | -0.64 | 0.02 |
| 3 | -0.98 | 0.03 | -0.98 | 0.03 |
| 4 | -0.69 | 0.02 | -0.69 | 0.02 |
| 5 | -0.03 | 0.02 | -0.03 | 0.02 |
| 6 | 0.03 | 0.02 | 0.03 | 0.02 |
| 7 | 0.07 | 0.02 | 0.07 | 0.02 |
| 8 | 0.52 | 0.02 | 0.52 | 0.02 |
| 9 | 0.34 | 0.02 | 0.34 | 0.02 |
| 10 | 0.15 | 0.02 | 0.15 | 0.02 |
| 11 | 0.56 | 0.02 | 0.56 | 0.02 |
| 12 | 0.12 | 0.02 | 0.12 | 0.02 |
| 13 | -1.08 | 0.03 | -1.09 | 0.03 |
| 14 | -0.43 | 0.02 | -0.44 | 0.02 |
| 15 | -0.53 | 0.02 | -0.53 | 0.02 |
| 16 | 0.49 | 0.02 | 0.49 | 0.02 |
| 17 | 0.21 | 0.02 | 0.21 | 0.02 |
| 18 | 0.46 | 0.02 | 0.46 | 0.02 |
| 19 | -0.49 | 0.02 | -0.49 | 0.02 |
| 20 | 0.30 | 0.02 | 0.30 | 0.02 |
| 21 | -0.36 | 0.02 | -0.36 | 0.02 |
| 22 | -0.43 | 0.02 | -0.43 | 0.02 |
| 23 | 0.23 | 0.02 | 0.24 | 0.02 |
| 24 | 0.63 | 0.03 | 0.63 | 0.02 |
| 25 | -0.25 | 0.02 | -0.25 | 0.02 |
| 26 | -0.95 | 0.03 | -0.95 | 0.03 |
| 27 | 0.27 | 0.02 | 0.27 | 0.02 |
| 28 | -0.59 | 0.03 | -0.59 | 0.03 |
| 29 | 0.42 | 0.02 | 0.42 | 0.02 |
| 30 | -0.07 | 0.02 | -0.07 | 0.02 |
| 31 | -0.64 | 0.02 | -0.64 | 0.03 |
| 32 | -1.09 | 0.03 | -1.23 | 0.03 |
| 33 | -0.26 | 0.02 | -0.35 | 0.03 |
| 34 | 0.07 | 0.02 | -0.02 | 0.03 |
| 35 | -0.22 | 0.02 | -0.37 | 0.03 |
| 36 | 0.20 | 0.02 | 0.12 | 0.03 |