Can Differential Rapid-Guessing Behavior Lead to Differential Item Functioning?

Christine E. DeMars

Steven L. Wise

James Madison University

(2007, April). Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Abstract

This investigation examined whether different rates of rapid guessing between groups could lead to detectable levels of differential item functioning (DIF) even in situations where the item parameters were the same for both groups. Two simulation studies were designed to explore this possibility. The groups in Study 1 were simulated to reflect differences between high-stakes and low-stakes conditions, with no rapid guessing in the high-stakes condition. Easy, discriminating items with high rates of rapid guessing by the low-stakes group were detected as showing DIF favoring the high-stakes group when using the Mantel-Haenszel index. The groups in Study 2 were simulated to reflect gender differences in rapid guessing on a low-stakes test. Both groups had some rapid guessing, but the focal group guessed more. Easy items with greater differences in rapid guessing were more likely to be detected as showing DIF. Our results suggest that there may be instances in which statistically-identified DIF is observed due to the behavioral characteristics of the studied subgroups rather than the content of the items.

Can Differential Rapid-Guessing Behavior Lead to Differential Item Functioning?

Unlike measurement in the physical sciences, achievement testing requires the cooperation of the objects of measurement (i.e., examinees). This means that whenever we administer an achievement test to an examinee, we implicitly assume that the examinee will try to do his or her best on the items. Despite our assumptions to the contrary, however, noneffortful responding by examinees is not uncommon. Consequently, because the item responses provide the basis for estimating the examinee's proficiency level, any responses that do not reflect that examinee's proficiency level will distort proficiency estimation, and thereby diminish the validity of the examinee's test score. Moreover, because noneffortful responses tend to be correct less frequently than those for which the examinee exhibits good effort, their presence tends to have a systematic negative bias on proficiency estimation.

Wise and Kong (2005) described three situations in which low examinee effort is likely to pose measurement problems. First, there are a number of low-stakes educational testing programs at the primary, secondary, and post-secondary levels for which test performance holds little or no consequences for individual examinees. Second, testing programs sometimes need to administer some of their test items in low-stakes settings. For example, new testing programs commonly field test items in "non-counting" testing situations prior to their being used in high-stakes operational testing, and the data from such administrations are often used in item calibration and/or test form construction. Third, a significant proportion of educational measurement research is conducted at colleges and universities in low-stakes settings using volunteer participants.

Examinee Effort and Response Time

When computer-based tests (CBTs) are used, item response time (i.e., the amount of elapsed time between the display of an item to an examinee, and the examinee's response) can be used to identify examinee response strategies. For example, it has been found that as time is running out during timed high-stakes tests, many examinees will switch response strategies from trying to work out the answers to remaining items (termed *solution behavior*) to quickly entering answers at random to those items (termed *rapid-guessing behavior*) in hopes of getting some correct (Schnipke, 1995a; Schnipke & Scrams, 1997).

More recently, Wise and Kong (2005) found that during untimed low-stakes tests, rapid-guessing behavior also occurs, and that in this context it can be attributed to low examinee effort. In addition, the evidence to date suggests that in low-stakes testing contexts, the amount of rapid-guessing behavior exhibited by examinees is unrelated to their proficiency levels as measured by criteria external to the test (Sundre & Wise, 2003; Wise & DeMars, 2005; Wise & Kong, 2005). This implies that item responses based on rapid guessing are generally uninformative for proficiency estimation on low-stakes tests.

Schnipke (Schnipke, 1995b; 1996; Schnipke & Scrams, 1997) noted that the peak of the item response time frequency distribution often occurred much sooner for incorrect responses than for correct responses. The response time distribution for the incorrect answers often had a sharp frequency spike occurring during the first few seconds (corresponding to rapid-guessing behavior), while the response time distribution for the correct answers had a broader distribution with a smaller peak (corresponding to solution behavior). Together, these distributions would form a bimodal distribution. Schnipke suggested classifying each item response as one of these behaviors using a threshold for each item that indicates the minimal amount of time for an examinee to be able to work out the answer to the item. Item responses occurring faster than this threshold would be classified as rapid guesses, with the remainder being deemed solution behaviors.

This classification of item responses has had several additional recent applications in low-stakes testing contexts. Wise and Kong (2005) developed a measure of the effort expended on a test by an examinee, termed response time effort (RTE), which was defined as the proportion of test items for which solution behavior had been exhibited. Wise (2006) introduced a corresponding measure of the effort received by a test item (termed response time fidelity, or RTF), which was defined as the proportion of examinees who exhibited solution behavior to that item. Wise and DeMars (2006) developed an effort-moderated IRT model, which specified for each item separate response functions for rapid-guessing behaviors and solution behaviors. Wise and DeMars showed that when rapid-guessing behavior was present in test data, the effort-moderated model showed better model fit and item parameter estimation, and yielded more valid scores than a traditional IRT model.

It should be noted that in this conceptualization, examinee test-taking effort is considered on an item-by-item basis. That is, an examinee's response strategies may (and in fact often will) vary across items. This change in strategies may be related to item characteristics; Wise (2006) found that items vary in the effort they receive, which suggests that here are some item features that influence the amount of effort exhibited by a given examinee. Wise found that item length and position were particularly salient features.

The Effort-Moderated IRT Model

The effort-moderated IRT model is based on the classification of each item response as either rapid guessing or solution behavior. For a given item *i*, there is a threshold, T_i, that represents the response time boundary between rapid-guessing behavior and solution behavior[1]. Given an examinee *j*'s response time, RT_{ij}, to item *i*, a dichotomous index of item solution behavior, SB_{ij}, is computed as

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

Wise and DeMars (2006) noted that the accuracy rates for responses under these strategies will typically be quite different, which implies that there is a different response function for each strategy. The effort-moderated model combines the two item response functions into a single model that is moderated by response strategy. The generic effort-moderated model for the probability of a correct response to an item using both types of response strategies can be expressed as:

$$P_i(\theta) = (SB_{ij})(\text{solution behavior model}) + (1 - SB_{ij})(\text{rapid} - \text{guessing behavior model}), \qquad (2)$$

with the dichotomous SB_{ij} defined as in Equation 1. Wise and DeMars investigated an effort-moderated model in which solution behavior was represented by the three-parameter logistic (3PL) IRT model, and rapid-guessing behavior was represented by a constant-probability model specified as $\mathrm{P_i}(\theta) = g_i$, where g_i is the reciprocal of the number of response options for item *i*. This model is shown in Equation 3.

$$P_i(\theta) = (SB_{ij})(\mathrm{c}_i + (1 - c_i)(\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}})) + (1 - SB_{ij})(g_i). \qquad (3)$$

Because SB_{ij} takes only the values 0 or 1 for a given item response, the effort-moderated model specifies two distinct response functions—one for a response based on solution behavior and one for a response based on rapid-guessing behavior. Depending on how long it took examinee *j* to answer item *i*, one or the other of the response functions is used to model the examinee’s response. Figure 1 shows an example of the two item response functions for a multiple-choice item with four response options. The response function for solution behavior is the familiar 3PL item characteristic curve (ICC), while the ICC for rapid guessing depicts a constant .25 probability of passing the item.

The general assumption that examinees uniformly give good effort to test items is tantamount to assuming that the ICC we conceptualize in IRT is equivalent to the solution behavior ICC. To the extent that rapid-guessing behavior is present, however, the functional relationship between proficiency and the probability of passing an item is distorted. For any proficiency value, the functional ICC is a weighted combination, or hybrid, of the solution behavior and rapid-guessing behavior ICCs, with the weights corresponding to the relative numbers of examinees exhibiting each type of behavior on the item. Figure 1 also shows two hybrid ICCs, for which the percentage the examinees exhibiting solution behavior was 80% and 90%, respectively. Note that the general effect of rapid guessing is to decrease the overall probability that the item is passed. Moreover, the greatest separation between the solution behavior ICC and the hybrid ICCs occurs at the upper end of the curve, suggesting that the distortion cause by rapid-guessing behavior has its greatest impact when an examinee encounters an item for which he or she would have a high probability of passing under solution behavior. However, note that when the *c* parameter is lower than the probability specified under rapid guessing (as in Figure 1), a low-proficiency examinee can actually have a slightly higher probability of passing the item under rapid-guessing behavior.

Rapid Guessing and DIF

The ICCs depicted in Figure 1 indicate that the proportion of examinees exhibiting rapid-guessing behavior moderates the functional relationship between proficiency and the probability of passing an item. This suggests that to the degree to which this proportion differs across examinee subgroups or measurement contexts, the functional ICCs will differ—which suggests differential item functioning (DIF). Thus, differential subgroup/context mean effort may represent an important source of DIF, though it does not appear to have been previously investigated.

It is not uncommon for DIF to be statistically identified, but uncorroborated by expert content review (Ercikan, Gierl, McCreith, Puhan, & Koh, 2004; Roussos & Stout, 1996; Shepard, Camilli, & Williams, 1984; Skaggs & Lissitz, 1992). If, under realistic conditions, differential effort can produce detectable DIF, then it represents an important, heretofore unexamined factor that may help explain the frequent incongruence between statistical identification of DIF and expert review.

There are several realistic scenarios under which effort-related DIF could occur. First, there may be examinee subgroups that exhibit different degrees of effort. For example, gender differences have been reported both in the rates of rapid-guessing behavior (Schnipke, 1995a; Wise, Kingsbury, Thomason, & Kong, 2004) and in self-reported test-taking motivation (Eklöf, in press). Second, sometimes items are pilot tested in a low-stakes context even though they will eventually be operationally administered in a high-stakes test. In such cases, there will be some rapid-guessing behavior due to low motivation expected under the low-stakes administration, but virtually none under the high-stakes administration. This could produce DIF across administration conditions.

The purpose of the present investigation was to examine the conditions under which differential examinee effort can produce DIF that is detectable using standard identification methods and criteria. Two studies were conducted. In Study 1, simulated examinee response time effort measures (RTEs) and item response time fidelity measures (RTFs) were based on those found for an actual test given under more motivating and less motivating conditions. The four item parameters of the effort-moderated model (*a*, *b*, *c*, & *g*) were randomly varied to assess their effects on the odds that an item would be flagged as showing DIF against one of the motivation groups given that the item parameters were invariant across the two groups. The effects of sample size were also studied. In Study 2, simulated examinee RTEs and item RTFs were empirically based on those found for male and female students in a low-stakes testing context. Based on the findings of Study 1, item parameters likely to lead to false DIF detection were used to simulate the data. The differences in RTE between gender groups were smaller than the differences between RTE for the more and less motivating conditions of Study 1; the purpose of Study 2 was to investigate whether smaller differences in RTE would still lead to flagging items as exhibiting DIF.

Study 1

The research question for Study 1 was: Can DIF be caused by groups that differ *only* in their rapid-guessing behavior? If so, which item characteristics (difficulty, etc.) make items most vulnerable to DIF in this situation?

Method

Empirical Scenario Used As a Basis for Simulation. For this study, the item parameters of the effort-moderated model were invariant across two groups, but one group was more highly motivated and showed no rapid-guessing behavior, while the other group had relatively high rates of rapid-guessing behavior. Thus, even with the same item parameters, the item trace lines would differ at least somewhat, as depicted earlier in Figure 1. The question was rather this difference would be large enough to detect through the Mantel-Haenszel DIF procedure. This question was prompted by an actual scenario in which test administration order had a large impact on rapid-guessing behavior, presumably due to lower motivation when the test was administered later in the sequence. In this assessment situation, the tests were used for program evaluation and students earned points toward a course grade by completing the tests, but the number of points did not depend on the test score; low scorers and high scorers earned equal credit so the tests were relatively low stakes. The test considered here was administered with seven other tests (two per week for four weeks). Observational evidence suggested that some students were becoming less motivated over time, rushing through the tests in later weeks, so the order of the tests was manipulated to check for order effects (see DeMars, in press, for a full description and results). In the more motivating context, the test studied here was administered in the first week; in the less motivating context it was administered in the last week. Though this manipulation of motivation may appear small, the differences in average RTE were striking. Mean RTE (proportion of test items for which solution behavior had been exhibited) was .999 in the more motivating condition and .845 in the less motivating condition (minimum = 0, maximum = 1). The distribution of RTEs for the less motivating condition is shown in Figure 2. The distribution was noticeably skewed, with a large proportion of examinees with RTEs near 1, similar to distributions from other low-stakes tests (Wise & Kong, 2005). The distribution from the more motivating condition is not shown because virtually all RTEs were 1. The RTFs (proportion of examinees who exhibited solution behavior) in the less motivating context ranged from .73 to .91 with a mean of .85 (the mean RTF must equal the mean RTE). The distribution of RTFs is displayed in Figure 3. Notice there were no items with RTFs of 1 in the less-motivating context; other studies have found RTFs of 1 or nearly 1 for items administered at the beginning of low-stakes tests (Wise, 2006), but the items in this study were administered in random order so there were no item position effects. These examinee RTEs and item RTFs from this earlier study were used in the present simulation study.

Data Simulation. In the data simulation, examinee RTEs and item RTFs were based on those found in the real data study described above. RTEs in the less motivating condition were distributed as shown in Figure 2, while all RTEs in the more motivating condition were equal to 1. The test had 30 items, with RTFs for the less motivated condition distributed as shown in

Figure 3. For the simulation, these 30 RTFs were replicated 500 times. To avoid having the same item parameters repeatedly paired with the same RTF each time, the four item parameters of the effort-moderated model were randomly varied across the 500 replications, creating 500 unique test forms with a total of 15,000 items. Item difficulties for each test form were randomly selected from a uniform distribution from -2.5 to 2.0, and item discriminations were randomly selected from a uniform distribution from 0.5 to 2.0. Uniform distributions were selected so that the more extreme items would be evenly represented. The less difficult items were expected to show the largest effects, so the range of difficulties included more negative items. Half of the *g*-parameters were set to .3 and half were set to .2; in previous work Wise (2006) found that rapid guessers were more likely to choose middle options such that for items with four options approximately 30% rapid-guessed the correct option when it was placed in the B or C position but only approximately 20% rapid-guessed correctly when the correct answer was in the A or D position. Half of the *c*-parameters were set to .15 because this value yielded good item fit in the real data, and the other *c*-parameters were set equal to the *g*-parameter. Conceptually, *c* would be lower than *g* if there were appealing distractors such that examinees using solution behavior who were uncertain of the correct answer would be drawn to a distractor more often than they would resort to guessing (Lord, 1974).

After item parameters and RTFs were simulated for 500 test forms of 30 items, θ’s and RTEs were generated. For each test form, 1000 focal group members and 1000 reference group members were simulated; this sample size was chosen to be large enough to estimate the effect size with good precision while small enough to be realistic. For the focal group, RTEs were randomly selected, with replacement, from the RTEs of the less-motivated group in the real data context described above. Because virtually all of the RTEs in the more-motivated group were 1, RTEs for the reference group members were set to 1. θ’s for both groups were drawn from a standard normal distribution. To simulate rapid guessing, for each examinee-by-item encounter the probability of rapid guessing was calculated as (1-RTE)(1-RTF)/(1-mean RTF). If a random number from the uniform [0,1] distribution was less than this probability, SB was coded 0; otherwise SB was coded 1. Equation 3 was then used to calculate the probability of a correct response, and another uniform random number was drawn to determine whether the item would be coded correct, if the random draw was less than the calculated probability, or incorrect if the random draw was greater than the probability.

Analyses. After the data were generated, the Mantel-Haenszel common-odds ratio α (Holland & Thayer, 1988) was calculated for each item. Within each test form, simulees were divided into score groups based on total score (0-30). The common-odds ratio was calculated as:

$$\hat{\alpha} = \frac{\sum_{j=0}^{30} R_{rj} W_{fj} / N_j}{\sum_{j=0}^{30} R_{fj} W_{rj} / N_j}, \qquad (4)$$

where R_{rj} is the number of reference group members in group *j* who gave the correct response, W_{fj} is the number of focal group members who gave an incorrect answer, R_{fj} is the number of focal group members who gave the correct response, W_{rj} is the number of reference group members who gave an incorrect answer, and N_j is the number of examinees in group j. In this context, the more-motivated group was the reference group, so the common-odds ratio represents the ratio of the more-motivated group’s odds of correct response to the less-motivated group’s odds of correct response, after conditioning on total score. A χ^2 test can be used to check whether the odds ratio is significantly different from 1.

The odds ratio was converted to the ETS Δ scale: $\Delta = -2.35 \ln(\alpha)$. When α is 1, Δ will be 0; testing whether the odds ratio is significantly different from 1 is equivalent to testing whether Δ is significantly different from 0. Odds ratios greater than one, favoring the reference group, correspond to negative Δ values and odds less than one correspond to positive Δ's. One advantage of the Δ's is that they are symmetrically distributed around zero while odds ratios are asymmetrically distributed around 1.

Items were flagged as exhibiting DIF favoring the more-motivated group if Δ was < -1 and significantly different from zero at *p* < .05. Items were flagged as exhibiting DIF favoring the less-motivated group if Δ was > 1 and significantly different from zero at *p* < .05. ETS procedure (Zieky, 1993) would classify these items as at least 'B' items, or moderate DIF.

Logistic regression was used to estimate the effects of RTF, *a*, *b*, *c*, and *g* on classifying items as DIF/no-DIF. These factors were treated as continuous predictors, except for *c* which was coded 1 if it was the same as *g* or 0 if it was fixed to .15. Each predictor was centered around its mean, so that each coefficient could be interpreted as the change in log-odds for each 1-unit increase in the predictor, holding the other predictors constant at their mean values. The focus was on effect size rather than statistical significance because with 15,000 items even very small effects could be statistically significant. $\mathrm{R}_{\mathrm{L}}^{2} = 1 - \left(-2\mathrm{LL}_{\mathrm{full}} / -2\mathrm{LL}_{\mathrm{intercept}} \right)$ (McFadden, 1974) was used to quantify the model effect. $\mathrm{R}_{\mathrm{L}}^{2}$ was recommended by Menard (2002) because it is not sensitive to the base rate (which was small in this study) or to sample size (which was quite large).

Results

Using the effect size and statistical significance criteria described above to flag moderate DIF, 18% of the items were flagged as favoring the more-motivated group and 1% were flagged as favoring the less motivated group. For predicting the log-odds of identifying an item as favoring the more-motivated group, the $\mathrm{R}_{\mathrm{L}}^{2}$ for the main effects model was .774. Each two-way interaction was entered individually in the model; each interaction increased R^{2} by less than .02

so no interactions were included in the final model. The largest interaction, *b* x *a* increased R^2 by .011. This interaction was interpretable; the compound effect of a low difficulty and a high discrimination was greater than the main effects of the two summed. However, given the small increase in R^2 this effect was not included in the final model. The estimates for the main effects model are displayed in Table 1.

Items with higher RTFs were less likely to be flagged. In other words, items on which few examiness rapid guessed were less likely to appear to show DIF. For each .10 increase in RTF, such as from .75 to .85, the log-odds decreased by 3.47. More discriminating items were more likely to be flagged. Each 1-unit increase in the slope, such as from 0.5 to 1.5 or 1 to 2, increased the log-odds by 3.36. More difficult items were less likely to be flagged. Each unit increase in difficulty, such as from -2.5 to -1.5, decreased the log-odds by 6.43. Items where rapid guesses were more likely to be correct were less likely to be flagged; the log-odds were 0.87 lower when *g* was .30 compared to .20. When the lower-asymptote equaled the rapid-guessing parameter, the log-odds were 0.55 greater than when the lower-asymptote was set to .15.

Probabilities are generally viewed as more interpretable than log-odds. Using the coefficients from Table 1, Figure 4 shows how the probability was influenced by item difficulty while holding the other predictors constant at their means. Figures 5-8 show how this relationship shifts for representative levels of each of the other predictors in turn. The probability is shown as a function of item difficulty in each figure because unless item difficulty was low the other predictors had little impact; if item difficulty were held constant at its mean very few items would be flagged regardless of the value of the other predictors so the probability function would be nearly flat. Conceptually, this is similar to an interaction effect, but recall that no interactions were included in the log-odds model. Due to the non-linear relationship between probability and log-odds, the effects of each of the other predictors on the probability depends on the level of item difficulty. Notice that the probability curves are essentially parallel until they come close to their upper or lower limits of 1 or 0, which translates to parallel slopes (no interaction) in the log-odds model.

A smaller proportion (1%) of items was flagged as favoring the less-motivated group. If *p*-values alone had been used to flag items this would have been expected by chance, but recall that the items were flagged based on the combination of statistical significance and the Δ effect size, so 1% was more than would be expected by chance[2]. This small percentage is likely not enough to cause practical concern, but the results are an interesting contrast. The $\mathrm{R}_{\mathrm{L}}^{2}$ for the main effects logistic regression model was only .158. Each of the two-way interactions explained less than 2% of the variance and these interactions were not included in the final model. The coefficients for the main-effects model are displayed in Table A1 in the Appendix. Items were more likely to be flagged when RTF was high; items with more rapid guessing were more likely to be flagged as showing DIF. For each .10 increase in RTF, the log-odds increased by 1.68. More discriminating items were again more likely to be flagged; for each unit increase in slope, the log-odds increased by 2.12. More difficult items were slightly more likely to show DIF, though the absolute value of the effect of item difficulty was much smaller than it was for predicting DIF favoring the more-motivated group. Setting *c* equal to *g* decreased the log-odds by 1.60. Higher *g* increased the odds of DIF. Again, because probabilities are more readily interpreted than log-odds, Figures A1-A3 in the Appendix illustrate the effects of RTF, *a*, and *b* in the probability metric. Because *c* and *g* each have only two levels, probabilities for each level are reported in Table A2 instead of graphed.

The probability of flagging an item as favoring the less-motivated group is clearly very small. Hence the practical implications, as will be discussed, are limited.

Discussion

More discriminating items were more likely to be flagged in either direction, as is commonly found in DIF studies[3]. For the other predictors, in contrast, the effects on the log-odds of flagging an item as favoring the more-motivated group were opposite the effects on the log-odds of flagging an item as favoring the less-motivated group. Items with lower RTFs were more likely to favor the more-motivated group and items with higher RTFs were more likely to favor the less-motivated group. When there are high rates of rapid guessing for an item, the proportion of students in the less-motivated group who give the correct response is lower than would be expected from their total scores. On the other hand, when there are low rates of rapid guessing, more examinees in the less-motivated group give the correct response than would be expected from their total scores. This is because the total score for examinees who frequently rapid guess tends to be lower than would be expected given θ; the total score is an underestimate of the ability that underlies their responses when they do use solution behavior.

Easier items were more likely to be flagged as favoring the more-motivated group, and more difficult items were slightly more likely to be flagged as favoring the less-motivated group. This would be predicted from Wise and DeMars’ (2006) findings that item difficulties are overestimated for easy items when examinees who used rapid-guessing behavior are included in the item parameter estimation. Easy items would be answered correctly by most examinees using solution behavior so low rates of correct response by rapid guessers have a particularly large impact on easy items. Referring back to Figure 1, easy items would be shifted to the left, so that more examinees would be in the range where there is a sizable difference between the hybrid and effortful trace lines.

The *c* and *g* parameters also influenced the odds of DIF flagging. When *g* was higher so that more rapid guessers would respond correctly, as would be the case when the correct response was a middle option, the item was less likely to favor the more-motivated group and more likely to favor the less-motivated group. When *c* was set to .15 instead of equal to *g* to simulate an item where one or more distractors were more attractive to low ability examinees than the correct response, the item was more likely to be flagged as favoring the less-motivated group because low-ability examinees were less likely to get the item right by using solution behavior than by rapid guessing. In other words, at the lower end of the score range the hybrid and effortful trace lines would cross, with the hybrid line above the effortful line.

It is important to re-emphasize that the proportion of items favoring the less-motivated group was very small. The results regarding these items were reported primarily to contrast with the results for items favoring the more-motivated group. The practical implications for the effects of RTE and the item parameters on items favoring the less-motivated group are probably small.

From these results we can conclude that differential rapid guessing can lead to DIF, particularly DIF favoring the group with less rapid guessing, when the groups differ in mean RTE by .15. This large difference has been observed when one group took the test first in a sequence of low-stakes tests and the other group took the same test last in the sequence. Similar differences might be found between high and low-stakes tests, such as an operational test and a pilot test. However, the groups in DIF studies are more often demographic groups, such as gender, race/ethnicity, or SES, not groups taking the test in different motivational contexts. In low-stakes settings, the mean RTE of these groups might differ, but to a lesser extent than the mean RTEs of groups in different testing contexts. In Study 2, the difference in mean RTE was based on observed gender differences within the less-motivated condition described in the introduction to Study 1; both groups included some rapid guessing, so the gender groups represented smaller differences in RTE than the test-order groups used in Study 1.

Study 2

The research question for Study 2 was: How large do the differences between groups in rapid-guessing behavior on an item have to be to cause DIF?

Method

Empirical Scenario Used As a Basis for Simulation. Again, while the data were simulated, a real-data context prompted the question and provided realistic values for the differences in rapid-guessing behavior. In the less-motivated context described in introducing Study 1, the mean RTF was .80 for men and .90 for women. Across the 30 items, the gender difference in RTF ranged from .05 to .16. These differences in RTF comprised the variable of focus in Study 2.

Data Simulation. As described above, the 30 items in the real data provided 30 RTF differences ranging from .05 to .16. These 30 RTF differences were replicated 500 times, with difference randomly matched to a different item difficulty each time. To avoid confounding the item RTF with the item RTF difference, the RTF averaged across the reference and focal group was kept constant at .85. For example, when the RTF difference was .05 the RTF for the reference group was .875 and the RTF for the focal group was .825, and when the RTF difference was .16 the RTF for the reference group was .93 and the RTF for the focal group was .78. Item difficulties were again randomly selected from a uniform distribution from -2.5 to 2.0 to simulate a broad-range test. To simplify interpretations of the primary variable of interest, difference in RTF, *a* was set to 1.25 for all items (high enough that DIF would be detectable, but low enough to be realistic), *g* was set to .25, and *c* was set to .15. Thus, there were two item factors in the design: difference in RTF and item difficulty.

RTEs for simulees were sampled with replacement from the real-data less-motivated condition, as described in Study 1 except that the RTEs were sampled separately from the male and female distributions (means of .80 and .90, respectively). For each of the 500 test forms, 1000 simulees were sampled for each group. The male group was treated as the focal group due to the higher level of rapid guessing. Two conditions were simulated for *impact*, or group differences in mean ability. In the *no-impact* condition, θs for both groups were drawn from N(0,1) distributions; the observed scores for the two groups differed only due to differences in RTE, not to differences in ability. In the *impact* condition, the θs for reference simulees were distributed N(0,1) and the θs for the focal simulees were distributed N(-0.5,1)[4]. SB was coded for each simulee by item encounter as described for Study 1, except that the mean RTF used to calculate the probability of rapid guessing varied by gender group. After RTEs, θs, and SBs were generated, the probability of correct response was calculated using the effort-moderated model in Equation 3. All item parameters were the same for reference and focal simulees.

Results

The Mantel-Haenszel Δ index, calculated as in Study 1, was used to flag items as exhibiting DIF if |Δ| was > 1 and significantly different from zero at p < .05. In the no impact condition, 8% of the items were flagged as favoring the reference group and 0.1% were flagged as favoring the focal group. In the group impact condition, only 3% of the items were flagged as favoring the reference group and again 0.1% were flagged as favoring the focal group. While 8% and 3% seem small, when random groups were formed with this data set only 0.06% of the items (9 out of 15,000) were flagged as favoring each group. Logistic regression was used to estimate the effects of *b* and difference in RTF on the log-odds that an item would be flagged for DIF favoring the reference group. Again, each predictor was centered around its mean, so that each coefficient could be interpreted as the change in log-odds for each 1-unit increase in the predictor, holding the other predictors constant at their mean values.

When there was no group impact, the R_L^2 was .625 for the main-effects model. As in Study 1, the interaction was statistically significant but increased R_L^2 by only .002 and was omitted from the model. A .10 increase in RTF difference increased the log-odds of flagging an item for DIF by 6.88 and a 1 unit increase in *b* decreased the probability by 4.29. In Figure 9, the meaning of this relationship is shown in the probability metric.

When the mean θ was 0.5 lower for the focal group, the *b* by RTF difference interaction effect was again very small (change in R_L^2 = .001) and was removed from the model. The R_L^2 was .524 for the main-effects model. A 0.10 increase in RTF difference increased the log-odds of flagging an item for DIF by 7.03 and a 1 unit increase in *b* decreased the probability by 2.87. Figure 10 displays this relationship translated from log-odds to probabilities. Holding *b* and RTF difference constant, an item was less likely to be flagged when there was group impact.

Discussion

The RTF differences used in the simulation were on average slightly smaller than in Study 1 (0.15 in Study 1 and 0.09 in Study 2), and, perhaps more importantly, both the reference and focal groups exhibited some rapid guessing. Fewer items were flagged for moderate or larger DIF. As would be predicted from Study 1, easier items and items with a greater difference in RTF were more likely to be flagged. The manifestation of DIF was less likely when the group mean θ's differed by 0.5. Because a focal group member would be less likely to get an item correct in this condition, even when using solution behavior, the presence of rapid guessing had less of an effect on the proportion correct. One way of visualizing this is to look back at Figure 1. The focal group would be shifted toward the lower end of the θ-axis, where the difference between the hybrid trace line and the effortful trace line is smaller.

Implications and Conclusions

When the focal and reference groups were modeled based on groups taking the test under less or more motivating conditions, 18% of the items showed DIF favoring the more-motivated group even though the item parameters were the same for both groups. If these conditions were to occur for a test given under pilot (less-motivating) and high-stakes operational (more motivating) conditions, easy items with lower RTFs in the focal group would appear to show parameter drift. This could impact pre-equating results. The effects, though, would be limited because only the easiest items would be affected.

Differential rapid guessing might also result in DIF for demographic groups that have different propensities to rapid guess. On the test used as a model for this study, in the less-motivating condition males rapid guessed more frequently than females. This led to detection of DIF favoring females for 8% of the items; these items had the largest differences in RTF and were the least difficult. Though this is not a large percentage of items, it could help explain some instances of empirically-identified DIF when there appears to be no corresponding content-related explanation. Also, this DIF could change if the amount of rapid guessing changed. For example, the groups might exhibit differential rapid guessing on a pilot test but have equally low (or zero) rates of rapid guessing on an operational test. Items that showed DIF in the pilot test due to differential motivation then would not show DIF in the operational test. This means that items might be unnecessarily discarded after pilot testing, a potentially costly consequence. Again, the percentage of items involved was relatively small, so the actual costs may be relatively small.

DIF was less frequent when the groups differed on θ. The focal group had lower mean θ and higher rates of rapid guessing. Previous work has shown that rapid guessing has little or no correlation with external measures of ability, such as SAT or course grades (DeMars, in press;

Wise & Kong, 2005). However, the θ for a particular test would be negatively associated with RTE if students with low RTEs generally put lower effort into items on which they used solution behavior. The θ that underlies item responses is a mixture of the student's knowledge and the student's willingness to exhibit that knowledge. Thus, if the groups differed at all on θ, the group with lower RTEs would tend to have lower average θ's (unless their average knowledge was enough higher that even with more rapid guessing their θ's remained higher than the other group). When lower RTE is combined with lower θ, DIF is less of a problem because there is less of a difference in the focal group between the proportion who get the item right through rapid guessing and the proportion who get the item right through solution behavior.

This study focused on rapid guessing on low-stakes power tests. Though not investigated here, the implications might generalize to high-stakes test which are slightly speeded due to time limits. Examinee subgroups may differ in the rates at which they tend to work on a timed high-stakes test (Llabre & Froman, 1987; O'Neill & Powers, 1993). When this is the case, then the test may be differentially speeded for the two groups, and differential amounts of strategic rapid-guessing behavior occur at the end of the test, which could potentially lead to DIF (Oshima, 1994).

One of the more puzzling aspects of previous DIF research has been the frequent failure of substantive analyses of item content/features to explain why some items exhibit statistical DIF. The results of the current study suggest an additional explanation for DIF—that it can be caused by differences in examinee effort. When trying to understand why a particular item exhibited DIF, measurement practitioners should therefore consider questions focused both on item content (i.e., what is it about the item's content that could have produced DIF?) and examinee behavior (i.e., why would one group have given less effort to this item than another group?). Thus, the results of this investigation have illustrated that item response time can play an important role in helping guide measurement practitioners to better understand the causes of DIF and whether or not item deletion is warranted when DIF occurs.

References

Chang, H.-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33,* 333-353.

DeMars, C. E. (in press). Changes in Rapid-Guessing Behavior over a Series of Assessments. *Educational Assessment*.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and Standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum Associates.

Eklöf, H. (in press). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing.*

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests, *Applied Measurement in Education*, *17*, 301-321.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kong, X. J., Wise, S. L., & Bhola, D. S. (in press). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*.

Llabre, M. M., & Froman, T. W. (1987). Allocation of time to test items: A study of ethnic differences. *Journal of Experimental Education*, *55,* 137-140.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39,* 247-264.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers of Econometrics* (pp. 105-142). New York: Academic Press.

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks: SAGE.

O'Neill, K. & Powers, D. E. (1993, April). *The performance of examinee subgroups on a computer-administered test of basic academic skills.* Paper presented at the meeting of the National Council on Measurement in Education, Atlanta, GA.

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

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24*, 293-322.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355-371.

Schnipke, D. L. (1995a). Assessing speededness in computer-based tests using item response times (Doctoral dissertation, Johns Hopkins University, 1995). *Dissertation Abstracts International, 57,* B759.

Schnipke, D. L. (1995b, April). *Assessing speededness in computer-based tests using item response times*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco. (ERIC Document Reproduction Service No. ED383742)

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

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

Shepard, L., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in intem bias research. *Journal of Educational Statistics, 9,* 93-128.

Skaggs, G., Lissitz, R. W. (1992). The consistency of detecting item bias across different test administrations: Implications of another failure. *Journal of Educational Measurement, 29,* 227-242.

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

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

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

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

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

Footnotes

[1] Kong, Wise, & Bhola (in press) showed that the various methods used for identifying item time thresholds tend to yield highly similar results.

[2] To verify the proportion due to chance, the simulees from both groups were randomly assigned to arbitrary “reference” and “focal” groups. With these random groups, 0.2% of the items were flagged as favoring one group or the other.

[3] Chang, Mazzeo & Roussos (1996) explained that, if the difference between item difficulties is constant, the area between the curves increases as the slope increases. For items that follow a 1 or 2PL model, $\Delta = 4a(b_R\text{-}b_F)$ (Donoghue, Holland, & Thayer, 1993). Thus, for a given difference in difficulty, more discriminating items have greater DIF. Roussos, Schnipke, and Pashley (1998) showed that the relationship is more complicated for 3PL items, but for easy items greater discrimination is related to greater DIF.

[4] This was more extreme than the group impact of -.27 standard deviations observed in the real data when scored by the effort-moderated model. It seems reasonable that any impact would favor the reference group, as it did in the real data, because the group with lower RTEs likely exerted lower effort even on the items to which they did not rapid guess and thus a lower theta-parameter would underlie their item responses.

Table 1

Log-odds of Favoring the More-Motivated Group

Factor	B	SE of B
Intercept	-9.353	0.245
RTF	-34.469	1.321
a	3.355	0.134
b	-6.425	0.167
g	-8.715	0.930
c[a]	0.546	0.091

[a] *c* was dummy-coded. After centering, -0.5 indicated *c* = .15, 0.5 indicated *c* = *g*

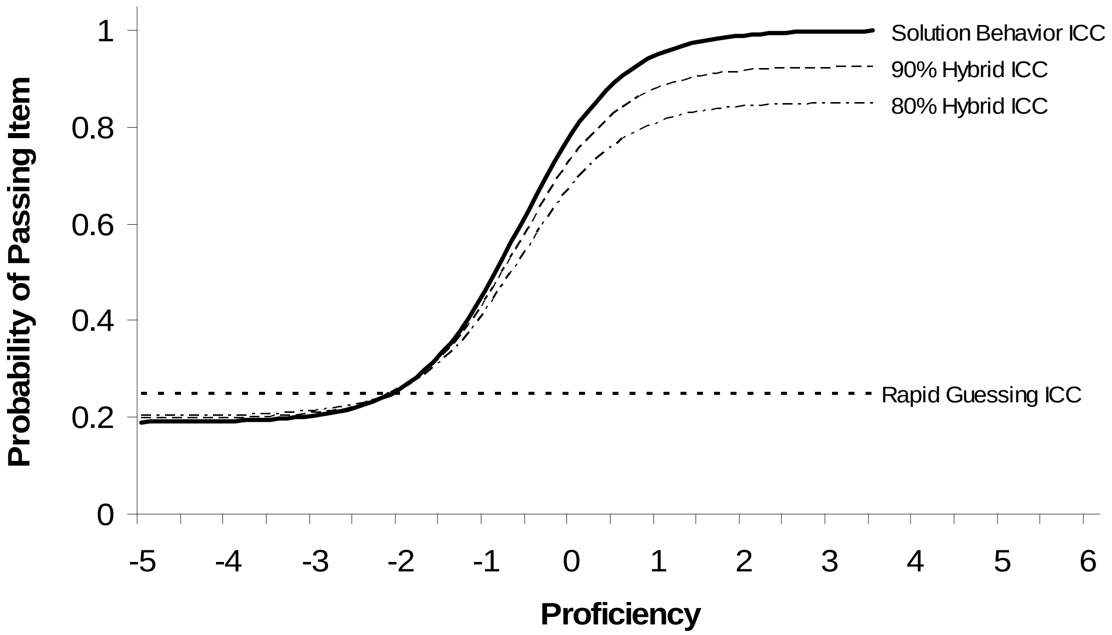

Figure 1. ICCs under solution behavior and rapid-guessing behavior, along with two hybrid ICCs.

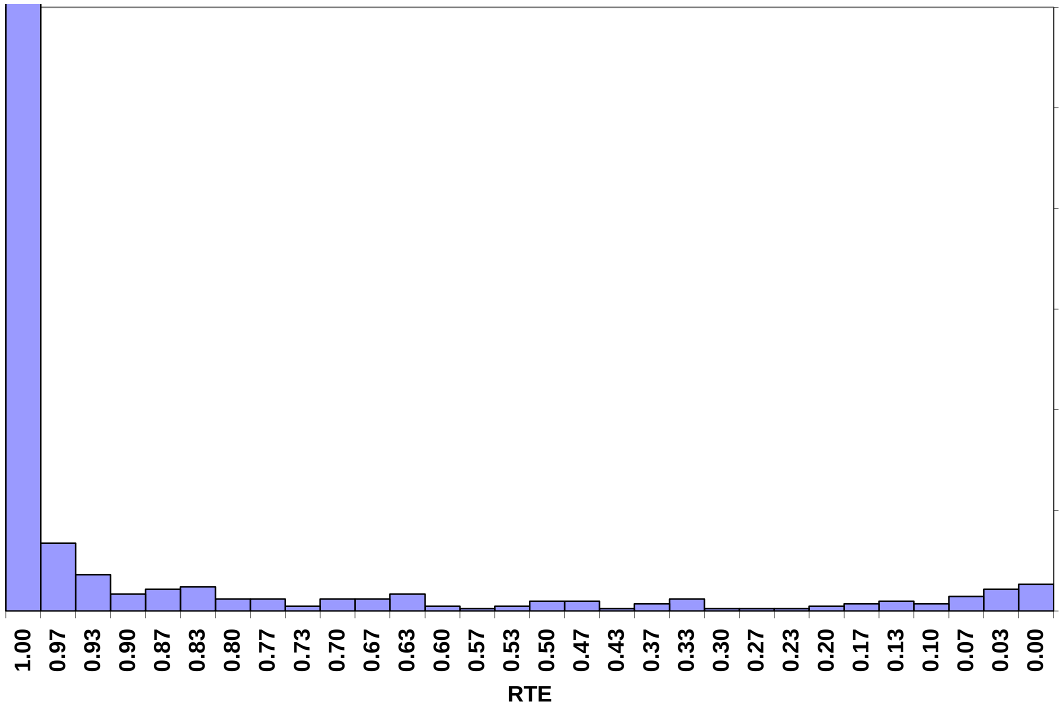

Figure 2: RTE distribution in the less motivating condition.

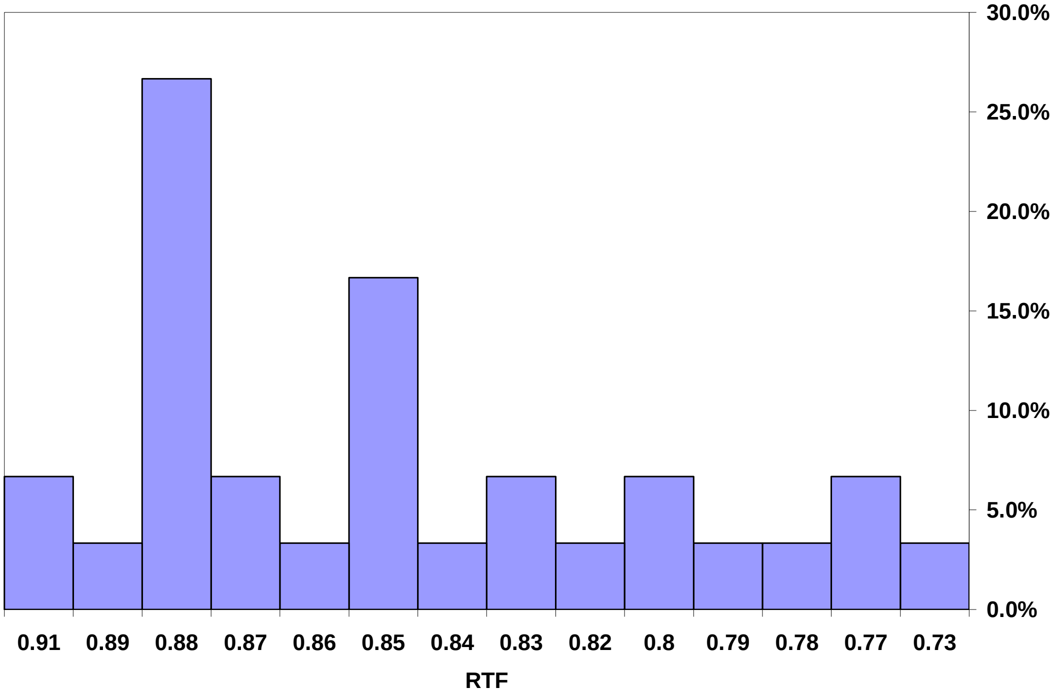

Figure 3. RTF distribution in the less motivating condition.

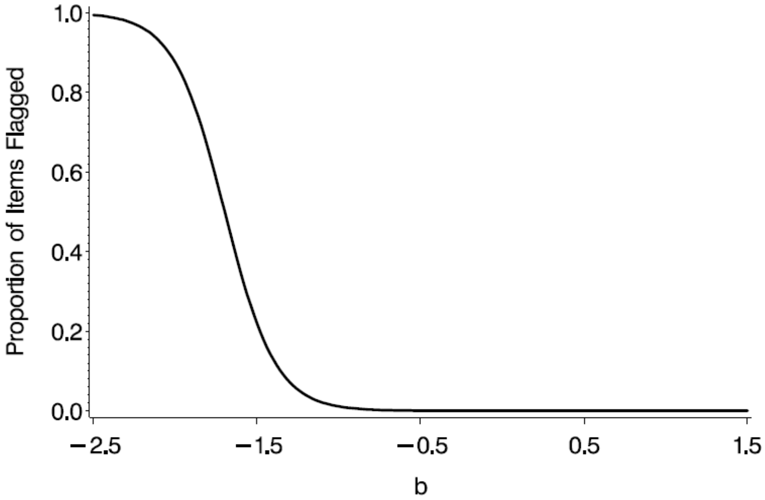

Figure 4. The effect of item difficulty (*b*) on the proportion of items flagged as favoring the more-motivated group, with other predictors held constant at their means.

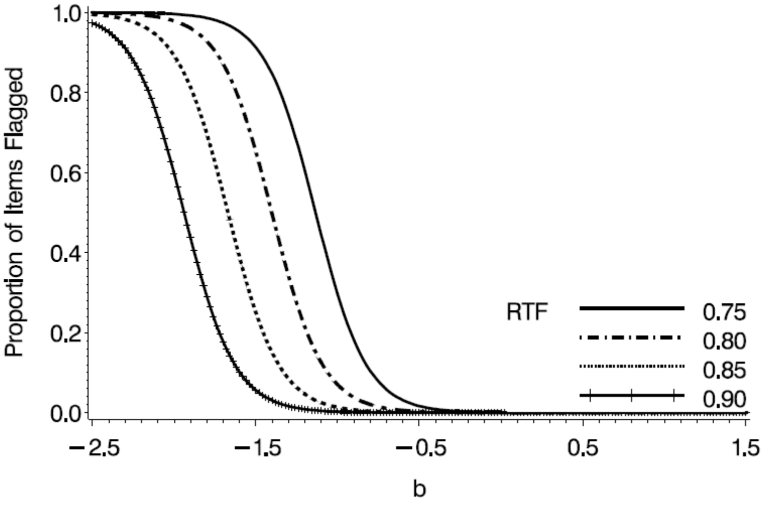

Figure 5. The effect of RTF on the proportion of items flagged as favoring the more-motivated group, across the range of item difficulty (*b*) with other predictors held constant at their means.

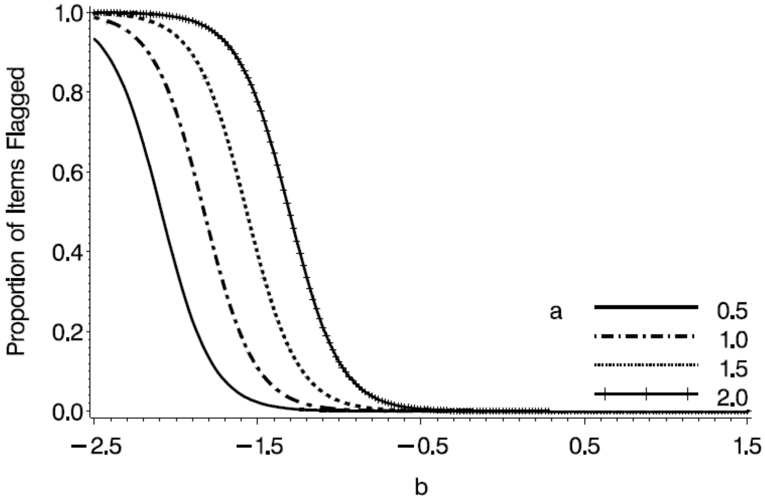

Figure 6. The effect of item discrimination (*a*) on the proportion of items flagged as favoring the more-motivated group, across the range of item difficulty (*b*) with other predictors held constant at their means.

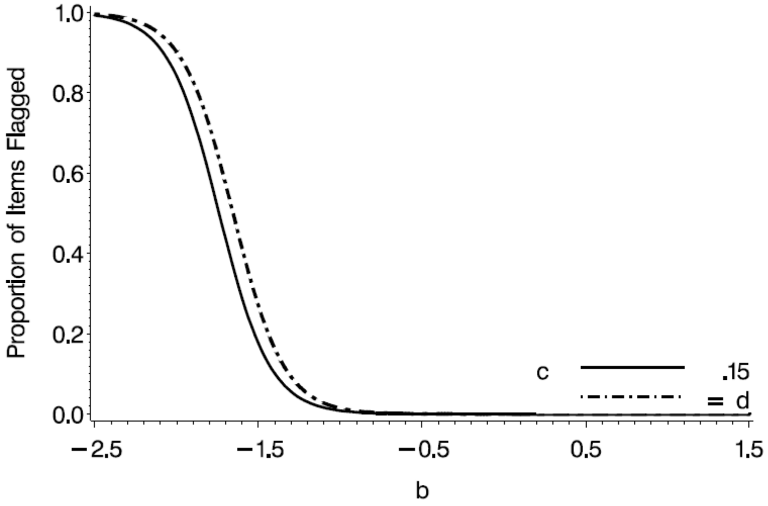

Figure 7. The effect of lower asymptote (*c*) on the proportion of items flagged as favoring the more-motivated group, across the range of item difficulty (*b*) with other predictors held constant at their means.

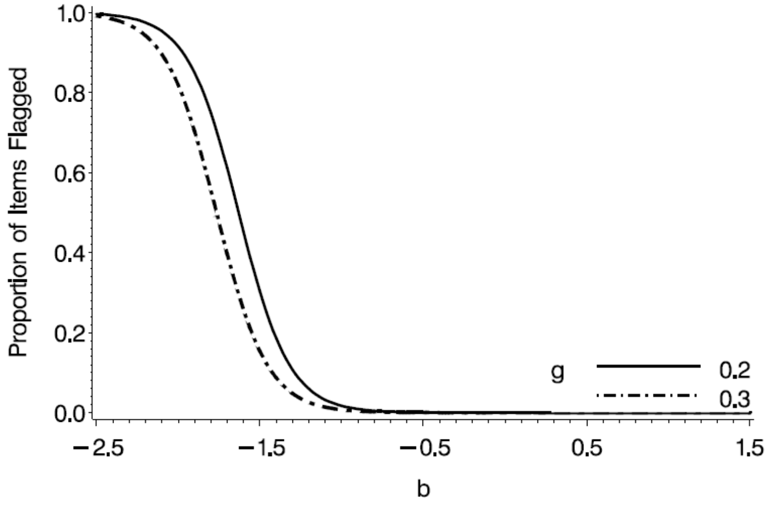

Figure 8. The effect of *g* on the proportion of items flagged as favoring the more-motivated group, across the range of item difficulty (*b*) with other predictors held constant at their means.

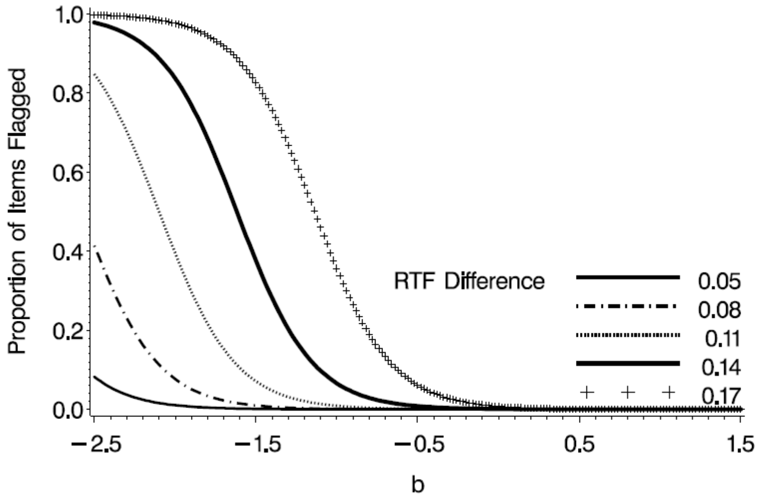

Figure 9. Relationship between reference and focal group RTF difference and probability of flagging an item for at least moderate DIF against the focal group, across the range of item difficulty (*b*), in the no group impact condition.

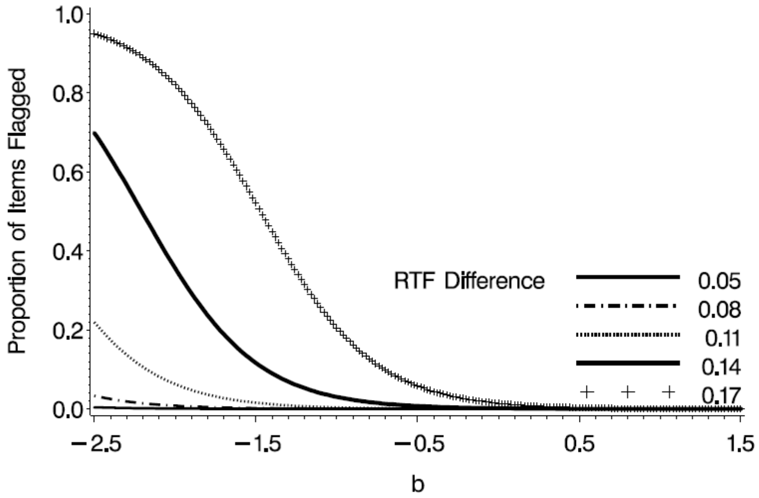

Figure 10. Relationship between reference and focal group RTF difference and probability of flagging an item for at least moderate DIF against the focal group, across the range of item difficulty (*b*), in the group impact = .5 condition.

Appendix

DIF Favoring the Less-Motivated Group

Table A1

Log-odds of Favoring the Less-Motivated Group

Factor	B	SE of B
Intercept	-5.628	0.156
RTF	16.796	2.683
a	2.124	0.235
b	0.383	0.068
c	-1.595	0.214
g	7.139	1.757

Table A2

Probability of Favoring the Less-Motivated Group

Factor and Level	Probability
c = .15	.008
c = *g*	.002
g = .20	.003
g = .30	.005

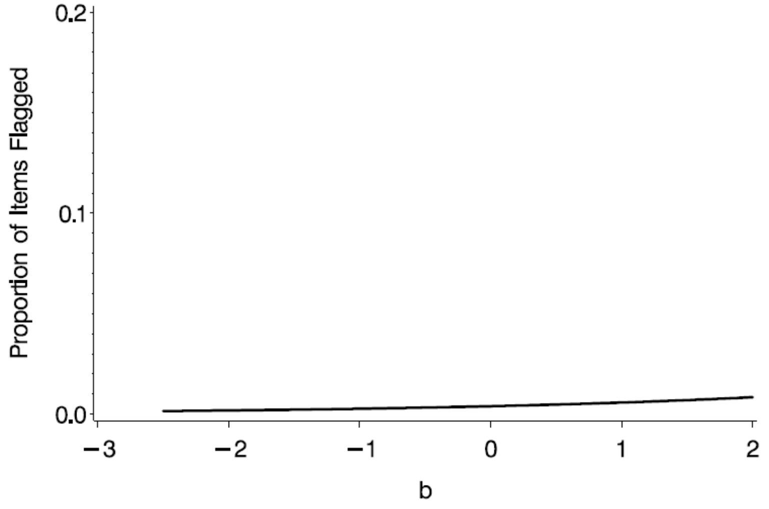

Figure A1. The effect of item difficulty (*b*) on the proportion of items flagged as favoring the less-motivated group, with other predictors held constant at their means.

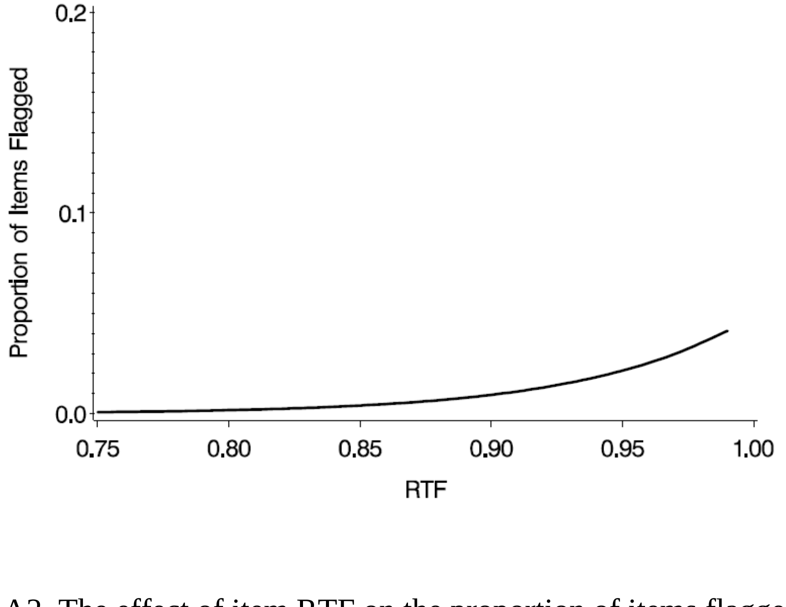

Figure A2. The effect of item RTF on the proportion of items flagged as favoring the less-motivated group, with other predictors held constant at their means.

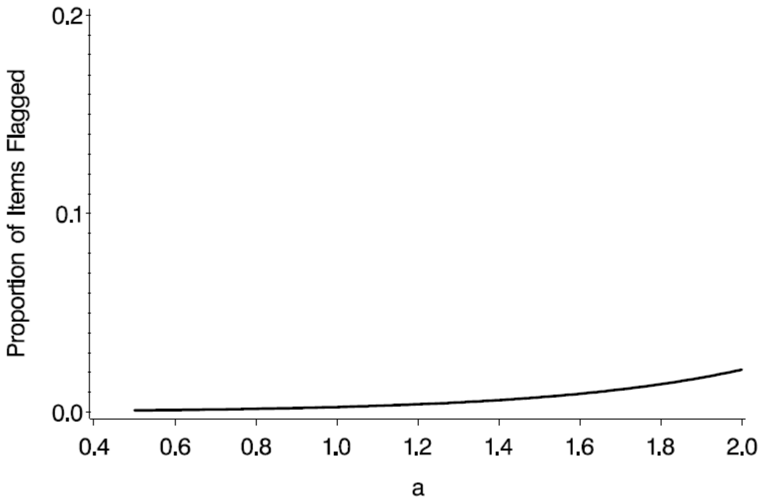

Figure A3. The effect of item discrimination (*a*) on the proportion of items flagged as favoring the less-motivated group, with other predictors held constant at their means.