The Interaction of Ability Differences and Guessing When Modeling Differential Item Functioning With the Rasch Model: Conventional and Tailored Calibration

Educational and Psychological
Measurement

2015, Vol. 75(4) 610–633

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164414554082

epm.sagepub.com

Christine E. DeMars¹ and Daniel P. Jurich²

Abstract

In educational testing, differential item functioning (DIF) statistics must be accurately estimated to ensure the appropriate items are flagged for inspection or removal. This study showed how using the Rasch model to estimate DIF may introduce considerable bias in the results when there are large group differences in ability (impact) and the data follow a three-parameter logistic model. With large group ability differences, difficult non-DIF items appeared to favor the focal group and easy non-DIF items appeared to favor the reference group. Correspondingly, the effect sizes for DIF items were biased. These effects were mitigated when data were coded as missing for item–examinee encounters in which the person measure was considerably lower than the item location. Explanation of these results is provided by illustrating how the item response function becomes differentially distorted by guessing depending on the groups' ability distributions. In terms of practical implications, results suggest that measurement practitioners should not trust the DIF estimates from the Rasch model when there is a large difference in ability and examinees are potentially able to answer items correctly by guessing, unless data from examinees poorly matched to the item difficulty are coded as missing.

Keywords

differential item functioning (DIF), Rasch, model fit

¹James Madison University, Harrisonburg, VA, USA

²National Board of Medical Examiners, Philadelphia, PA, USA

Corresponding Author:

Christine E. DeMars, Center for Assessment and Research Studies, James Madison University, MSC 6806, Harrisonburg, VA 22807, USA.

Email: demarsce@jmu.edu

Differential item functioning (DIF) occurs when item parameters vary across groups, after adjusting for any differences in the groups' mean abilities. The DIF may be because of construct-irrelevant variance, so including DIF items in the scores may be unfair. Thus, it is important to accurately identify DIF items. However, when there is a nonzero lower asymptote in the item response function (IRF), modeling DIF with a two-parameter logistic (2PL) or linear logistic regression model can lead to inflated Type I error rates if there is a large difference in the focal and reference groups' mean ability (DeMars, 2010). Several factors can introduce a nonzero lower asymptote into the response function. The nonzero lower asymptote may be due to correct guessing because none of the distractors consistently lure low-ability examinees away from the correct answer, or examinees may have some level of knowledge of the correct answer due to factors beyond the primary ability. For brevity, the nonzero lower asymptote will be referred to in this study simply as guessing. The inflated Type I error results from differential distortion in the response function. To accommodate the lower asymptote, the slope of the function will flatten. The degree to which it flattens depends on the relative difficulty of the item; the more difficult the item, relative to the group mean ability, the more the slope will flatten (Wells & Bolt, 2008; Yen, 1981).

In addition to the distortion in the IRF, if a latent model that does not include a lower asymptote is applied, the latent scale itself will become distorted (Yen, 1986). Specifically, the units at the lower end of the scale will be stretched out. This scale distortion occurs to allow the model to provide better fit for the more difficult items where guessing is a problem. However, it may also bias the estimates of the discrimination parameters for the easier items that would have been more accurately estimated without the scale distortion because there is less correct guessing on the easier items. Both the IRF distortion and scale distortion vary depending on the ability distribution, leading to false detection of DIF when the ability distributions of the reference and focal groups are not equal.

Use of Rasch modeling for DIF detection in the presence of group differences in mean ability when there is correct guessing has not been studied as thoroughly. This presents an important gap in the literature given the frequent application of Rasch models in high-stakes educational testing, where DIF analyses are a necessary component of validation. The Rasch model differs from the 2PL in that it does not allow individual items to vary in the degree to which their slopes flatten to match the lower asymptote. Thus, the Rasch model does not allow for DIF in the item discriminations. Because DIF cannot manifest in the discrimination parameters, it is possible that the false DIF introduced by not accounting for the nonzero lower asymptote will appear in the difficulty estimates instead. This property also makes it more difficult to analytically derive the results.

One procedure to reduce the effects of guessing on parameter recovery in the Rasch model is to code responses as missing when the initial estimate of the item difficulty is considerably higher than the initial estimate of the person ability (Andrich & Marais, 2014; Andrich, Marais, & Humphry, 2012; Choppin, 1983), a procedure

labeled *tailored* calibration. The rationale is that this removes the responses most likely to be affected by guessing. Although it is unknown whether an individual response is guessed or not, coding a valid response as missing should not bias parameter estimates as long as the data are missing at random (MAR; Rubin, 1976). MAR means that after conditioning on the covariates, the missingness is unrelated to the parameter of interest. In this context, the covariates are the other item responses and the parameters being estimated are the item difficulty and person ability. To strictly meet MAR, the response under consideration for coding as missing should be omitted from the initial estimate of person ability and item difficulty. However, as each response is only a small part of the parameter estimate, the violation of MAR is likely to be mild.

Choppin (1983) coded data as missing whenever an examinee's probability of correct response, conditional on the initial ability estimate, was $<.25$, slightly above random guessing on 5-option items. Choppin illustrated the procedure, but did not give details on its accuracy, noting only that "initial results support the suggestion that it can lead to substantial improvement in measurement on 'difficult' multiple-choice tests" (pp. 32-33).

Similarly, Andrich et al. (2012) coded responses as missing when the modeled probability of correct response was $<.30$ on a test where the simulated probability for very-low-ability examinees was $.14$. They termed this model the *tailored* model because items were targeted to better match examinee ability. They simulated data such that very-low-ability examinees guessed correctly with probability of $1/7$, and the probability of correct guessing decreased with increasing ability. The item locations based on the tailored model were essentially unbiased, although standard errors were larger for more difficult items because there were fewer nonmissing responses. Conversely, when all responses were treated as valid, if the mean item difficulty was constrained to zero the locations of easy items were biased upward (appeared more difficult) and the locations of difficult items were biased downward (appeared easier). When the mean location of the easiest items was set equal across the two models, the negative bias in the other item locations increased with increasing difficulty. The same procedure was followed by Andrich and Marais (2014), who showed that the bias in item parameters produced corresponding bias in person estimates if likely guesses were not recoded as missing. High-ability examinees were penalized because the difficulty of the most difficult items was underestimated.

Another approach is to code responses as wrong when the probability of correct response is low (Waller, 1989). Although coding data as missing will not bias the parameter estimates, assuming the missingness meets the MAR assumption, coding responses as incorrect could potentially bias the parameter estimates, if responses that were true corrects (not guesses) were miscoded as wrong. Waller's ability removing random guess (ARRG) model was a 2PL model with responses coded wrong if the probability was less than a selected value. The cutoff value was selected based on model fit, but never exceeded the inverse of the number of options. The ARRG fit better than the 3PL or conventional 2PL for only 20% of real data sets

studied (Waller, 1989). Weitzman (1996) used a similar assumption in correcting the proportion correct for each item and each examinee, and then analyzing the proportion correct values with the Rasch model. Although Weitzman's procedure did not involve coding individual responses, it still required stronger assumptions than coding responses as missing. Given the strong assumption under this coding scheme that examinees should have responded incorrectly to these items, this study only explores tailored calibration when selected responses are coded as missing, rather than incorrect.

Using conventional calibration, where all responses are treated as valid, bias in item difficulty estimates may differ depending on the group's mean ability, leading to the appearance of DIF in non-DIF items. If tailored calibration reduces the bias in item difficulty estimates, it should reduce the group differences in bias and thus reduce the appearance of spurious DIF. Even a small reduction in item parameter bias can improve accuracy of DIF statistics considerably. For example, suppose that the difficulty for item *Z* is underestimated by 0.7 logits in the lower scoring group but only by 0.2 logits in the higher scoring group, using conventional calibration. If tailored calibration reduces the underestimation to 0.2 and 0.1, respectively, the estimated DIF will be reduced from 0.5 logits to 0.1 logits.

This study examines the effects of group ability differences, item difficulty, and item discrimination on Type I error and power for DIF modeled by the Rasch model in the presence of correct guessing, using both conventional and tailored calibration. In the following sections, data are simulated to show these effects. The purpose of the simulation is to demonstrate these effects empirically, before explaining the results analytically in the next section. Although the simulation is not necessary to understanding the implications of conducting DIF studies with the Rasch model for 3PL data when the group abilities differ, it helps make the explanation more concrete by illustrating the effects empirically.

Method

Data were simulated to follow a 3PL model:

$$P(\theta) = c + (1 - c) \frac{e^{1.7a(\theta - b)}}{1 + e^{1.7a(\theta - b)}} \quad (1)$$

where $P(\theta)$ indicates the probability of correct response given θ , the examinee's ability or level of the construct measured, and the item parameters (more fully expressed as $P[x = 1|\theta, a, b, c]$), a indicates the item discrimination, b indicates the item difficulty, and c indicates a lower asymptote. The c parameter is the probability of correct response for examinees with very low θ , often referred to as a pseudoguessing parameter because guessing is one reason that the low- θ examinees may have a nonzero probability of correct response. For the Rasch model, $c = 0$, the 1.7 is omitted from the model, and $a = 1$; the item discrimination is transferred to the variance of θ such that a given group of examinees will have greater variance for a more discriminating

set of items. In Rasch formulations, ability is commonly symbolized by β instead of θ , and item difficulty is symbolized by δ instead of b . For consistency, the notation in Equation (1) will be used throughout this exposition. Typically, the size of the scale units is set in the 3PL model by fixing the variance of θ to 1 in a reference group, so that a 1-unit change in θ represents a 1-standard deviation difference. In the Rasch model, a 1-unit change in θ represents a difference of 1 logit; the log-odds of correct response increases by 1 when θ increases by 1. If $c = 0$ in the 3PL, a 1-unit change in $\theta = 1.7a$ logits. As such, the logit difference varies from item to item depending on a . If $c \neq 0$, the logit difference is also a function of θ . Thus, comparisons are not invariant under the 3PL model.

Two hypothetical test forms were generated for this empirical demonstration. Each test form had 45 items, with $c = .2$, and b evenly spaced from -2.1 to 2.1 , and with three items at each value of b . For Form 1 all $a = 1$, and for Form 2 $a = 0.6, 1.0, 1.4$. In the DIF condition, one moderately easy ($b = -1.5$), one middle ($b = 0$), and one moderately difficult ($b = 1.5$) item favored the reference group, and three items with corresponding difficulties ($-1.5, 0, 1.5$) favored the focal group. The difference in bs was calculated to produce a Δ difference of $|1.75|$ (logit difference = 0.745). The Δ difference is a DIF effect size frequently used at Educational Testing Service to evaluate the practical significance of DIF (Zieky, 1993), where $\Delta = 2.35 * \text{logit difference}$. An item with $|\Delta \text{ difference}| > 1.0$ and statistically significantly different from 0 is classified as a “B” item, showing moderate DIF. An item with $|\Delta \text{ difference}| > 1.5$ and significantly different from 1 is classified as a “C” item, showing large DIF. Negative differences favor the reference group and positive differences favor the focal group. In the 3PL model, as noted, the logit difference varies as a function of θ . The average logit difference was calculated by integrating the difference over the θ distribution, weighting by the density of the focal and reference groups combined. The integration was approximated by evaluating the difference at 49 points from $\theta = -4$ to $\theta = 4$. The difference in b corresponding to Δ difference = 1.75 ranged from 0.363 (easy, high-discrimination item) to 1.387 (difficult, low-discrimination item).

The mean ability difference, often termed *impact*, was set to 0 or 1. When the mean ability difference was 0, the reference and focal group’s ability distributions were equivalent at $\theta \sim N(0, 1)$. When the mean ability difference was 1, 0.5 was added to the reference group mean and subtracted from the focal group mean, to produce higher average ability in the reference group. Sample size was 1,000 per group. For each condition, 500 replications were conducted. Responses generated for each replication were based on probabilities computed using Equation (1); if a random draw from a uniform distribution was lower than the model-implied probability, the response was coded as a correct answer.

Items were calibrated using WinSteps 3.8 (Linacre, 2013a). Both groups were calibrated concurrently, with the DIF table requested. This is the method recommended by Linacre (2013b; Section 19.24) to keep the scale equivalent for both groups. First, item difficulties and examinee abilities were estimated for both groups

pooled together. Then, the person abilities were held constant and item difficulties were estimated for each group.¹ The difference in b was transformed to the Δ scale by multiplying by 2.35.

Two calibrations were conducted for each data set. The *conventional* calibration used all item responses, except those few from examinees with perfect or zero scores. The *tailored* calibration used the Winsteps option “cutlo = 1” so that examinee by item encounters in which the preliminary person measure was more than 1 logit below the preliminary item measure were effectively coded as missing before the final item calibration.² One logit below corresponds to a probability of .27, comparable to Andrich et al.’s (2012) choice of .3 or Choppin’s (1983) .25.

Simulation Results and Discussion

Item Fit

As a preliminary step, item fit was assessed, as should be done whenever calibrating items. Most Rasch software, including WinSteps, provides weighted indices of item fit, Infit, and unweighted indices, Outfit. Outfit is more sensitive to misfitting responses from persons far from the item difficulty. The magnitude of the DIF is quantified by the Infit or Outfit mean-square (MS). To create a statistical significance test, the MS can be transformed to a z statistic using the Wilson–Hilferty cube root transformation. Although there has been much discussion among Rasch analysts about using MS versus statistical significance (e.g., see Linacre, 2003; Smith, Schumacker, & Bush, 1998), Wilson (2005) has recommended following the procedure commonly used in statistics: First test for statistical significance, then interpret the MS as an effect size if the significance test is rejected. For dichotomously scored items, Wright and Linacre (1994) recommended MS between 0.8 and 1.2 for high-stakes tests or 0.7 to 1.3 for “run of the mill” tests. They also suggested that MS in the range of 0.5 to 1.5 are “productive for measurement,” and values between 1.5 and 2.0 are “unproductive for construction of measurement, but not degrading.”

Because of the large sample size, the focus was on interpreting the magnitude of the misfit, not the statistical significance. For the conventional calibration, although there was a clear pattern of increasing misfit with increasing item difficulty,³ the average magnitude of the Infit MS was within Wright and Linacre’s guidelines for all items. Within individual replications, Infit MS often slightly exceeded 1.2 for the most difficult items. Outfit was more problematic for the conventional calibration; easy items tended to have $MS < 1$, often less than the recommended 0.8, and difficult items tended to have $MS > 1$, often greater than 1.2.

Fit was better for the tailored calibration, especially for the most difficult items. None of the items had Infit MS outside the range of 0.8 to 1.2 in any of the replications. The mean Infit MS for the most difficult item was 1.06, compared with 1.20 for conventional calibration. Improvements for Outfit were even larger; the mean Outfit MS for the most difficult item was 1.07, compared with 1.53 for conventional calibration. For the easiest items, where little or no data were coded missing, the Outfit still

improved somewhat because the better estimation of the more difficult items produced less scale distortion. The mean Outfit MS for the easiest item was 0.79, compared with 0.64 for conventional calibration.

Based on the average Infit MS across replications, most items would be retained for both the conventional and tailored calibrations. Based on the Outfit values, analysts might choose to drop the easiest and most difficult items for the conventional calibration, although the values seldom fell into the “degrading” (Linacre, 2002) range so the decision is not clear cut. For illustrative purposes, all items were retained for the DIF analysis.

Differential Item Functioning Estimates

Figures 1 and 2 show the mean of the estimated Δ -differences for the items where there was no DIF simulated. The Δ -differences were calculated from the estimated b -differences. First, the constant a test form is shown in Figure 1. With no mean ability difference, the Δ -differences were unbiased in both calibration methods. When there was a large mean ability difference, difficult items tended to falsely appear to favor the focal group. To a somewhat lesser extent, easy items tended to favor the reference group. Importantly, the absolute value of the bias was far lower for the tailored calibration. For the conventional calibration, mean Δ values for the hardest items extended into the “B” range using the Educational Testing Service classification system. In contrast, the tailored calibration mean Δ values stayed well within the A range (minimal DIF).

Next, the varying a test form is shown in Figure 2. As with the constant a condition, there was no bias in the Δ -differences when there was no mean ability difference, regardless of item difficulty or discrimination. When mean ability difference = 1 standard deviation, for the most difficult items, the bias was similar to the bias observed in the corresponding constant a condition. Using conventional calibration, for the easiest items there was almost no bias when a was low. However, there was large bias in the Δ -differences when a was high. With the highest a , the easy items appeared to favor the reference group to an even greater extent than the difficult items appeared to favor the focal group. In this condition, the easiest items had mean Δ values in the “C” range. Using tailored calibration, less discriminating items appeared to favor the focal group regardless of item difficulty. Easy items with moderate or high a -parameters appeared to favor the reference group, but to a lesser degree than with conventional calibration. Importantly, the estimated Δ -difference was nearly unbiased for the difficult items with moderate or high a -parameters when using tailored calibration.

Given the variance of the Δ estimates across replications, for the conventional calibration many of the Δ -differences would extend into the “C” range even when the mean was in the “B” range. To illustrate this, empirical 90% confidence intervals are shown in Figure 3 for the easiest and hardest items, Test Form 1, when mean ability difference = 1. This figure also illustrates one possible drawback to using the

Figure 1. Estimated Δ -difference for the non-differential item functioning items, conditional on item difficulty, for the constant item discrimination ($a = 1$) test form. The conventional calibration is in the upper panel and the tailored calibration is in the lower panel.

tailored calibration: The confidence intervals for the most difficult items were somewhat larger for the tailored calibration because so many item responses were coded missing. For the most difficult item, 65% of the responses were coded missing when

Figure 2. Estimated Δ -difference for the non-differential item functioning items, conditional on item difficulty, for items of varying discriminations.

there was no mean ability difference. When the mean ability difference was 1, 46% of the responses from the reference group and 80% of the responses from the focal group were coded missing for the most difficult item. In operational testing, this

Figure 3. Empirical 90% confidence intervals for the Δ -difference for the easiest and most difficult non-differential item functioning items when the mean ability difference = 1 and the item discriminations were constant.

Figure 4. Estimated Δ -difference for the differential item functioning items, conditional on item difficulty, for the constant item discrimination ($\alpha = 1$) condition. Vertical bars extend 1 standard error in each direction (approximately 68% confidence intervals).

Figure 5. Estimated Δ -difference for the differential item functioning items, conditional on item difficulty, for varying item discrimination. Vertical bars extend 1 standard error in each direction (approximately 68% confidence intervals).

would typically still leave thousands of responses. It might be more problematic if small samples were used for field testing.

For the power study, Figures 4 and 5 illustrate the mean and standard error of the estimated Δ -differences for the six DIF items. The true Δ -difference was generated to be $|1.75|$. In the constant a test form (Figure 4) with no mean ability difference, the absolute value of Δ was nearly unbiased for the easy and middle difficulty items, but overestimated for the difficult items. The estimated Δ -difference appeared to be more biased using tailored calibration, but in some sense the parameter estimated was different under tailored calibration. The true Δ -difference was defined as the mean across the entire ability range, but the tailored calibration focused on the portion of the ability range 1 logit below the item location and higher. Because the log-odds difference increases as θ increases, due to decreased guessing, the magnitude of the DIF was greater than 1.75 in this range. To keep a constant average Δ -difference of 1.75, the data had to be generated with increasingly larger b -differences for harder items. If the true Δ -difference were defined as the log-odds difference that would occur if there were no correct guessing ($1.7 * 2.35 * [b_{ref} - b_{foc}]$) instead of as the average log-odds difference, then the tailored calibration would underestimate the absolute value of the parameter less than the conventional calibration.

When there was a large mean ability difference, effect sizes for difficult items were positively biased using conventional calibration, which meant they were underestimated in absolute value for items that favored the reference group and overestimated for items that favored the focal group. The converse was true for easy items. Using tailored calibration, the estimated DIF effect sizes were more similar to the corresponding estimates with no mean ability difference.

In the varying a test form (Figure 5) with no mean ability difference, the Δ was nearly unbiased for the easy, low a item but was negatively biased for the easy, high- a item (less so for the tailored calibration). For the medium difficulty items, Δ showed some negative bias. In absolute value, the Δ for the item favoring the reference group was inflated and the Δ for the item favoring the focal group was deflated. The difficult items showed bias of a similar magnitude to that in the constant a conditions. As noted, the tailored calibration Δ -differences were larger because they disregarded the low end of the ability range, where the DIF is smaller.

When there was a mean ability difference, the easy low- a item's Δ was less biased than in the constant a condition, using conventional calibration, but biased in the opposite direction using tailored calibration. However, the easy high- a item's Δ was more negatively biased than in the constant a condition, even appearing to slightly favor the wrong group using conventional calibrations. The difficult items' Δ s were more similar to those in the constant a condition.

To summarize the results using conventional calibration: When there was no mean ability difference, the Δ -difference was estimated accurately for non-DIF items, but was underestimated (in absolute value) for easy DIF items and overestimated (in absolute value) for hard DIF items. For easy items, the bias was minimal when the a was low. For difficult items the a had little effect. When there was a large mean ability difference, easy non-DIF items appeared to favor the reference group, and correspondingly Δ -differences were negatively biased for easy DIF items and were positively biased for difficult DIF items. Negative bias indicates that the absolute value of the Δ -difference was overestimated for items that favored the reference group but underestimated for items that favored the reference group. Results clearly show the ability difference influenced how DIF manifested. To clarify why these results occurred and why tailored calibration helped reduce the impact of ability differences, the next section will discuss how the 3PL response function was distorted when the Rasch model was applied.

Distortion of the Response Functions

To help explain these results, the following section elaborates on the impact of mean ability differences on the IRFs under DIF situations. As mentioned, when the 2PL model is applied to 3PL data, the response function will become distorted to better fit the data. Specifically, the slope will flatten to better meet the responses in the guessing range (DeMars, 2010; Yen, 1981). This effect increases with the difficulty of the item, relative to the distribution of θ . For difficult items or low-ability groups, the a -

Figure 6. Estimated item response function, by calibration method, when the mean ability difference = 1, for an easy item (true $b = -1.8$) and a difficult item (true $b = 1.8$). The thin solid line represents the data, and the dashed and dotted lines show the best fit for the reference and focal groups, respectively.

parameter is negatively biased. The b -parameter also tends to be negatively biased, because the items appear to be easier than they really are. In the Rasch model, the slope of the response function is estimated for the set of test items as a whole,⁴ and the b -parameter is estimated conditional on this slope such that the resulting function best fits the data. When the mean ability is equal for both groups, the estimated response functions for non-DIF items will be similar, but when the group means are unequal each group's estimated response function adjusts differently. Figure 6 shows the data function, and the best-fitting Rasch IRF, for the second-easiest (true $b = -1.8$) and second-hardest ($b = 1.8$) items, when mean ability difference = 1. When minimizing differences between the data and the estimated response function, the differences are weighted by the sample density. Because the focal group has lower ability, the low ability levels are weighted more for the focal group, and the high ability levels are weighted more for the reference group. Thus, for the easy item, the focal group estimate of b is pulled up more to try to match the data in the low-ability region. For the difficult item, the reference group has a higher estimated b to better match the data in the high-ability region and the focal group has a lower estimated b to better match the data in the low-ability region. Consequently, both items exhibit DIF, and the DIF is in opposite directions for hard and easy items.

Figure 6 also illustrates the effects of item discrimination. When the discrimination was low (left panel), the IRF for the easy item could be fit well with the Rasch model, so there was little difference between the focal and reference item difficulties. But the IRF for the high-discrimination (right panel) easy item was harder to fit, and thus weighting by the ability density led to different estimates of the item difficulty for the two groups. For the hard items, neither the low nor the high discrimination items fit the Rasch model, and each led to differential difficulty estimates. For the hard items, the difference was similar across discrimination levels. For the tailored calibration, the data in the lower ability range was effectively removed for the hardest item. The focal group function was not disproportionately influenced by the data at the low end of the θ range as much as it was using conventional calibration. As a result, the focal group IRF was not more biased than the reference group IRF, so there was little bias in the Δ -difference. Particularly for the less discriminating item, it is interesting to see that neither group matched the data very well, but the focal group IRF could not be pulled further away than the reference group as it was in the conventional calibration.

Person Estimates

When many items are affected by guessing, the measurement scale itself also becomes distorted. The units in the Rasch metric stretch to account for the fact that the log-odds (logit) of correct response change slowly in the region where the probability of correct response is just above the lower asymptote. Or equivalently one could say that the Rasch metric compresses the 3PL units. The relative unit sizes change to better match the test characteristic curve. In terms of log-odds, the 3PL metric is not of equal interval; a one-unit change in θ at the low end of the scale does not yield the same logit difference as a one-unit change in θ at the high end of the scale. Stretching and compressing the metric helps make each θ unit closer to a logit, although it can never eliminate the disparity in the units. The degree of metric distortion depends on the mix of item difficulties and the θ distribution.

Figure 7 shows the average scale recovered. Variation in the item discrimination and mean ability difference had little impact on the scale so the figure is not repeated for each condition. To make the scales more commensurate, the Rasch scale was rescaled such that the mean person measure (analogous to θ) was zero and the adjusted variance (true score variance) was equal to the generating variance.⁵ At the low end of the scale, the values are spaced increasingly further apart in the conventional Rasch model, and at the high end they are pushed closer together. As a result, as the item location or ability moves away from the center, the estimates become positively biased. Additionally, the locations of the 0 points are not perfectly aligned because an examinee at the mean on the 3PL scale would be slightly below the mean on the Rasch scale due to skewness introduced by the scale distortion. For the tailored calibration, however, the distortion is minimal at the high end and considerably

Figure 7. Estimated scales for the conventional and tailored Rasch calibration, compared with the three-parameter logistic (3PL) scale used to generate the data.

reduced at the low end. In the range from -2 and above, the metric of the tailored calibration matches the 3PL metric well.

This distortion in the θ estimates under conventional calibration provides additional insight into the underestimation of the absolute value of the DIF effect for easy items and the overestimation of the absolute value of the DIF effect for difficult items. At the easy end of the scale, 1.75 units on the 3PL scale corresponds to fewer units on the conventional Rasch scale, and conversely at the difficult end of the scale. This does not explain the appearance of DIF in the non-DIF items when there is a large mean difference in θ . For that, the discussion in the previous section is more relevant. Notice also that there is little distortion in the center of the scales, where both group means were located. In the group mean difference conditions, when averaged across replications, the difference between the means was within 0.006 of the generating difference of 1.

Effects of Other Study Design Factors

To keep the illustration simple, the only factors that varied were item difficulty, item discrimination, and mean ability difference. Some additional factors will be briefly considered here: 2PL data, unbalanced DIF, and sample size.

Two-Parameter Logistic Data

The interpretation of the interaction between item difficulty and item discrimination was necessarily confounded by the presence of more guessing for harder items. As a

Figure 8. Estimated Δ difference, conditional on item difficulty, for the two-parameter logistic (2PL) data.

further exploration of this interaction, Form 3 was simulated with the same parameters as Form 2, except that $c = 0$.

For the non-DIF items, mean Δ estimates were near zero for all items when there was no difference in mean ability (Figure 8) because item difficulties were equally biased for both groups. But when there was a large difference in mean ability, all the less discriminating items appeared to favor the focal group and all the more discriminating items appeared to favor the reference group. For the conventional calibration, these effects were slightly accentuated for the easiest and most difficult items. This occurred because the difficulty was more extreme for one group, and as the difficulty became more extreme the estimate of the item difficulty became biased further away from the mean for high- a items but further in toward the mean for low- a items. The bs for easy, low- a items were positively biased, but more so for the reference group because they were relatively easier; thus, the bs were higher for the reference group and appeared to favor the focal group. The bs for difficult, low- a items were negatively biased, but more so for the focal group because they were relatively harder; thus, the bs were again higher for the reference group and appeared to favor the focal group. The converse was true for the high- a items, where the bias was negative for the easy items (more so for the reference group) but positive for the difficult items (more so for the focal group). Greater negative bias (or smaller positive bias) for the reference group means the item appears to be easier for the reference group. The

Table 1. Estimated Δ for Differential Item Functioning Items, Two-Parameter-Logistic Model.

Item: Group favored	Conventional calibration		Tailored calibration	
	Mean Δ	SE of Δ	Mean Δ	SE of Δ
$\theta_R \sim N(0, 1) \theta_F \sim N(0, 1)$				
Easy, low- a : R	-2.05	.33	-2.03	.34
Easy, high- a : F	1.50	.36	1.60	.38
Medium, low- a : R	-2.10	.28	-1.99	.29
Medium, high- a : F	1.41	.24	1.55	.27
Difficult, high- a : R	-1.42	.33	-1.66	.64
Difficult, low- a : F	2.13	.31	1.97	.40
$\theta_R \sim N(0.5, 1) \theta_F \sim N(-0.5, 1)$				
Easy, low- a : R	-1.21	.33	-1.26	.34
Easy, high- a : F	0.59	.36	0.83	.38
Medium, low- a : R	-1.31	.28	-1.47	.30
Medium, high- a : F	0.59	.23	0.94	.27
Difficult, high- a : R	-2.32	.42	-1.96	.80
Difficult, low- a : F	2.94	.32	2.24	.52

Note. R indicates that the item difficulty is lower for the reference group, biased against the focal group; F indicates that the difficulty is lower for the focal group, biased against the reference group.

mean Δ estimates were $\leq |1|$, so most of the DIF would be considered small, on average. However, within some replications the Δ estimate would extend into the B range for some items, especially with small sample sizes where the standard error of Δ would be greatest.

Tailored calibration was used in the main study to address the issue of guessing, which was not present in the data for Form 3, but it also influenced the trends described for the large mean ability difference condition. For the easy items, the bias in the Δ difference was similar to that using conventional calibration, but the absolute value of the bias decreased as difficulty increased. A narrower range of ability was considered in fitting the more difficult items, so the differential bias in difficulty estimates caused by the variance in discrimination was less.⁶

For the DIF items (Table 1), when the mean ability was equal in both groups, the magnitude of the DIF was overestimated in absolute value for the low- a items and underestimated in absolute value for the high- a items. When the group means differed, this effect was added to the effects seen for the non-DIF items. For the difficult items, the absolute value of the bias was greater using conventional calibration, although of course the standard errors were larger using tailored calibration.⁷ The amount of data coded missing varied with the item discrimination because the initial difficulty estimates for the hardest items increased with discrimination. When the mean ability difference was 1, 72% of the responses from the reference group and 94% of the responses from the focal group were coded missing for the most difficult item, least discriminating item. The values were 93% and 99% for the most difficult

item, most discriminating item. These values were greater than for the 3PL data because the item difficulty estimates were higher without guessing.

Unbalanced Differential Item Functioning

To minimize interactions between and confounding of conditions, in the examples above the DIF favoring the reference group was balanced by DIF favoring the focal group (balanced for Form 1, as nearly balanced as practical for Form 2). Unbalanced DIF is also possible. As an extreme example of unbalanced DIF, all six DIF items were resimulated to favor the reference group, and the parameters were re-estimated without using any purification procedure. As a result, the θ s for the focal group were underestimated; when a focal group member and a reference group member had the same estimated θ , on average the focal group member had a higher true θ . When both groups had the same ability distribution, this meant that the non-DIF items appeared to slightly favor the focal group ($\Delta \approx 0.2 - 0.3$). When the focal group had the lower mean ability, with focal group $\sim N(-0.5, 1)$ and reference group $\sim N(0.5, 1)$, the non-DIF items that appeared to favor the reference group when the DIF was balanced did so to a lesser extent. The non-DIF items that appeared to favor the focal group when the DIF was balanced did so to a greater extent when the DIF was unbalanced. Again, the effect of group ability differences was much smaller using tailored calibration.

Without applying any purification, the estimated Δ -differences for the DIF items increased slightly by the same $0.2 - 0.3$ as the non-DIF items. Because the Δ differences were negative, this meant they became smaller in absolute value, which was sometimes an increase in the absolute value of the bias and sometimes a decrease, depending on the direction of the bias with balanced DIF earlier in Figure 4. The pattern was the same for both calibration methods.

Often, DIF studies use purification, where either DIF items are identified in an earlier round and removed from the anchor in further calibrations (Candell & Drasgow, 1988; Hidalgo-Montesinos & Lopez-Pina, 2002; Miller & Oshima, 1992; Raju, van der Linden, & Fleer, 1995), or a small number of items that initially show the least DIF are used as anchors in the next calibration (Wang, 2008; Woods, 2009), or both methods are combined (Wang, Shih, & Sun, 2012). Otherwise, if the DIF is unbalanced, using all items in the anchor can lead to false detection of non-DIF items as favoring whichever group is disfavored by the majority of DIF items (Andrich & Hagquist, 2012; Woods, Cai, & Wang, 2013). Depending on the purification method and criterion for choosing anchor items, some of the non-DIF items thus might be mistakenly removed and some DIF items might be left in the anchor set, particularly when using conventional calibration, potentially affecting the DIF detection of other items. Purification procedures are beyond the scope of this study, but the complexities of unbalanced DIF should be acknowledged.

Sample Size

Sample size typically has no effect on bias, but the standard errors of the estimates should increase as sample size decreases. This expected pattern was generally true for the Δ -difference estimates. The data used in the main study, where there were 1,000 examinees per group in each replication, were divided into replications with 500 per group or 100 per group. Sample size had little effect on the mean Δ -difference, conditional on true item difficulty (Figures 1 and 2), except that for the very easiest items the mean Δ -differences were slightly more negative (by about 0.25) for the small sample than the medium or large sample when the mean ability difference was 1, using either calibration method. This was because the distribution of the Δ -difference estimates was negatively skewed for the small sample for the easiest items. The empirical standard error of the Δ -difference was of course larger for the small sample. More of the effect sizes for the non-DIF items would fall into the “C” range for the small sample using conventional calibration, although fewer would meet the statistical significance criterion. Similarly, for the DIF items, sample size did not influence the mean bias, except for the easiest items; as noted for the non-DIF items, the smallest sample size had a negatively skewed Δ -difference distribution when the mean ability difference was 1, which pulled down the mean slightly for the easy item that favored the reference group.

Discussion and Implications

In a Rasch DIF study where the group mean ability difference is large and the data include correct guessing, the estimated response functions will differ by group. This can lead to the appearance of DIF for non-DIF items and bias in the effect sizes (with corresponding decrease or increase in power) for the DIF items. When the group means are equal, however, the Δ -difference effect size should be unbiased for non-DIF items and have relatively small bias for DIF items, even using conventional calibration.

The problem with distorting the response function arises when groups have different abilities, or equivalently when an item is calibrated with other groups of items of different average difficulty. Then the same item will have different difficulty estimates, depending on the group of people or items. The Rasch model with conventional calibration should not be used for DIF studies where the group ability difference is large and correct guessing seems plausible.

The issues introduced by the effects of guessing combined with ability differences could be eliminated by using the 3PL model to account for guessing. However, strategies for mitigating these effects within the Rasch framework are also important as operational contracts for psychometric work may require the Rasch model, or may have a long history of Rasch model use making changing models politically unfeasible. In addition, practitioners may wish to retain the desirable properties of the Rasch model, including its parsimony and the property of specific objectivity. Using tailored calibration, treating responses likely to be guesses as missing data allows for

mitigating the problem of guessing while continuing to use the Rasch model. Of note, under conventional calibration, the most difficult items tend to have the largest residuals from the estimated response function. In practice, these difficult items cannot all be discarded due to the increasing need to measure rigorous standards and have accurate estimates of high-ability examinees. However, it is necessary to consider that tailored calibration will reduce the sample size used in estimating DIF, thus increasing the standard error of the estimates. The reduction in sample size or the increase in standard error should be monitored prior to making inferences from tailored calibration estimates. In operational testing with many thousands of examinees, the reduction in sample size may not be a problem in terms of the standard error of the item parameter estimates. Apart from the statistical justification and the empirical results supporting tailored calibration, conceptually it may be justified by a comparison with computerized adaptive testing (Andrich & Marais, 2014). When a test is delivered through computerized adaptive testing, few responses are available for low-proficiency examinees on difficult items. With tailored testing, the data are tailored after test completion rather than on the fly.

But conceptually and politically, large losses of data may be questionable. Additionally, ideally DIF would be detected in field tryouts, which may involve only a few thousand students and much smaller representation of some studied groups. Furthermore, the sample sizes of lower ability groups are disproportionately affected by tailored calibration because more group members are in the guessing range. This problem may be exasperated by rigorous testing programs with many items that exceed the proficiency levels of many students.

Illusory DIF in the presence of large group mean differences is also seen for observed score-based DIF procedures, such as the Mantel-Haenszel, but for different reasons. In the Mantel-Haenszel procedure, the odds ratio is estimated conditional on observed score and averaged over the score distribution. If the scores are not perfectly reliable and the group means are unequal, examinees matched on observed score will not be matched on true score unless the data follow a Rasch model (Meredith & Millsap, 1992; Uttaro & Millsap, 1994; Zwick, 1990). Because of this mismatch, when there is no DIF, more discriminating items will appear to favor the reference group and less discriminating items will appear to favor the focal group (Li, Brooks, & Johanson, 2012; Uttaro & Millsap, 1994; Zwick, 1990). This mismatch accounts for a small portion of the false DIF seen with the Rasch model as well, because all examinees with the same observed score have the same estimated θ but different mean values of true θ when the model does not fit. Importantly, increasing the reliability of the scores will minimize the mismatch of true scores conditional on observed scores and thus minimize the false DIF due to varying item discrimination for the observed-score procedures. Increasing the reliability will have little effect on the false DIF using the Rasch model because most of the false DIF is due to a different cause. Additionally, the effects of item difficulty on DIF are in the same direction, but smaller, for observed score DIF procedures. If there is correct guessing, more difficult items appear to favor the focal group (Uttaro & Millsap, 1994; Wainer & Skorupski,

2005) because more examinees are in the guessing range, where there is little difference between the response functions.

Although alternative statistical models are likely the least expensive way to address the problem, a potentially better safeguard against the false DIF situation described in this study lies in test development. If the test includes multiple-choice items, careful attention should be given to the distractors. If distractors can be developed to capture the common misconceptions of lower ability students (Briggs, Alonzo, Schwab, & Wilson, 2006; Hermann-Abell & DeBoer, 2011) or to at least seem plausible (Haladyna, 2004, p. 120), correct guessing can be minimized because low-ability students will tend to select a distractor rather than the correct answer. Although this will not necessarily make the item discriminations more equal, it should at least help the lower asymptote to approach zero and limit the possibility of false DIF.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. This procedure is only effective when the DIF is approximately balanced between the two groups. If a preponderance of the DIF items favors one group, the non-DIF items will appear to favor the other group because the abilities will be biased. In such contexts, a subset of items should be selected as anchor items, as discussed later.
2. As described earlier, MAR is not theoretically perfectly met using the cutlo procedure. To check if the small violation of MAR made a practical difference, an additional data set was created where MAR was met, by definition. For one replication, the item and person parameters were calibrated 90,000 times, each time omitting one item and one person. Then, when coding a response for person j as missing for item i , ability was estimated from the item calibration without person j and without item i . Thus, the missingness for any item was conditioned only on the other item responses. The calibration using the MAR data yielded slightly less biased DIF estimates than the automated cutlo procedure, but bias was small for both calibrations.
3. Details of the fit analysis are available on request.
4. Typically, the a -parameter is fixed and the standard deviation of θ is free; a larger standard deviation of θ corresponds to a greater slope in the response function.
5. The multiplicative constant was between 0.96 and 0.98 for the conventional calibration, so the units remained almost equivalent to logits. The multiplicative constant was approximately 0.74 for the tailored calibration. For this scale adjustment, only the targeted

- responses used in the tailored item calibration were used in the estimation of the person measures.
6. Coding data as missing when the preliminary estimate of $\theta - b$ was above 1, as well as below -1 , can decrease the bias for the easier items. Although not shown in figure, this procedure yielded estimated Δ differences < 0.5 in absolute value for the non-DIF items, regardless of item difficulty. Ignoring data further from the item difficulty in both directions helps produce more accurate parameter estimates because when an item's discrimination is lower than average, as θ moves away from the item difficulty the residuals from the IRF based on the average slope are increasingly larger than the residuals from the IRF based on the item-specific slope, and conversely for items of higher than average discrimination. The difference between the IRFs is smaller near the item difficulty.
 7. Again, coding responses as missing when $\theta - b > 1$ further decreased the bias because of group ability differences for the easy and middle difficulty items, but increased the standard errors. When the group means differed, 93% of the reference group data and 87% of the focal group data were coded missing for the items with $b = -2.1$.

References

- Andrich, D., & Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics, 37*, 387-416.
- Andrich, D., & Marais, I. (2014). Person proficiency estimates in the dichotomous Rasch model when random guessing is removed from difficulty estimates of multiple choice items. *Applied Psychological Measurement, 38*, 432-449.
- Andrich, D., Marais, I., & Humphry, S. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics, 37*, 417-442.
- Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*, 33-63.
- Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*, 253-260.
- Choppin, B. (1983). *A two-parameter latent trait model* (CSE Rep. No. 197). Center for the Study of Evaluation, University of California, Los Angeles. Retrieved from <http://www.cse.ucla.edu/products/reports/R197.pdf>
- DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement, 70*, 961-972.
- Haladyna, T. M. (2004). *Developing and validating multiple choice items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Hermann-Abell, C. F., & DeBoer, G. E. (2011). Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. *Chemistry Education Research and Practice, 12*, 184-192.
- Hidalgo-Montesinos, M. D., & Lopez-Pina, J. A. (2002). Two-stage equating in differential item functioning detection under the graded response model with the Raju area measures and the Lord statistic. *Educational and Psychological Measurement, 62*, 32-44.
- Li, Y., Brooks, G. P., & Johanson, G. A. (2012). Item discrimination and type I error in the detection of differential item functioning. *Educational and Psychological Measurement, 72*, 847-861.

- Linacre, J. M. (2002). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3, 85-106.
- Linacre, J. M. (2003). Rasch power analysis: Size vs. significance: Standardized chi-square fit statistic. *Rasch Measurement Transactions*, 17, 918.
- Linacre, J. M. (2013a). Winsteps® (Version 3.80.0) [Computer software]. Beaverton, OR: Winsteps.com.
- Linacre, J. M. (2013b). *A user's guide to WINSTEPS, MINISTEP Rasch-Model computer programs*. Retrieved from <http://www.winsteps.com/>
- Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, 57, 289-311.
- Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement*, 16, 381-388.
- Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, 19, 353-368.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, 2, 66-78.
- Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement*, 18, 15-25.
- Wainer, H., & Skorupski, W. P. (2005). Was it ethnic and social-class bias or statistical artifact? Logical and empirical evidence against Freedle's method for reestimating SAT scores. *Chance*, 18, 17-24.
- Waller, M. I. (1989). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement*, 13, 233-243.
- Wang, W.-C. (2008). Assessment of differential item functioning. *Journal of Applied Measurement*, 9, 387-408.
- Wang, W.-C., Shih, C.-L., & Sun, G.-W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement*, 72, 687-708.
- Weitzman, R. A. (1996). The Rasch model plus guessing. *Educational and Psychological Measurement*, 56, 779-790.
- Wells, C. S., & Bolt, D. M. (2008). Investigation of a nonparametric procedure for assessing goodness-of-fit in item response theory. *Applied Measurement in Education*, 21, 22-40.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.
- Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, 33, 42-57.
- Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, 73, 532-547.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8, 370.
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245-262.

- Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, 23, 299-326.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-348). Hillsdale, NJ: Lawrence Erlbaum.
- Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics*, 15, 185-197.