Psychological Test and Assessment Modeling, Volume 61, 2019 (2), 227-257

Relationship between item characteristics and detection of Differential Item Functioning under the MIMIC model

Daniella A. Rebouças[1] *& Ying Cheng*[2]

Abstract

Differential item functioning (DIF) occurs when individuals of the same true latent ability or psychological trait from different demographic populations are found to have different chances of endorsing an item category. The ability to identify such items depends on many factors, including the sample size of each demographic group, average true latent trait scores in each group, the chosen DIF assessment method, the magnitude of DIF effect and the quality of the anchor set. An anchor is a group of items free of DIF that establish a common metric between groups. If the anchor is contaminated, that is, if it contains a DIF item, the common metric is inappropriate. The current literature rarely addresses the relationship between item parameters, anchor selection, and subsequent DIF detection. In this two-part study, we show that the power of DIF detection is high when the anchor has highly discriminating items. Additionally, DIF items of large discrimination and moderate difficulty have generally high power when using a correctly specified anchor, given a fixed DIF effect size. Implications for anchor selection and DIF effect size research are discussed.

Keywords: differential item functioning, anchor, item difficulty, item discrimination, effect size

[1] University of Notre Dame

[2] *Correspondence concerning this article should be addressed to:* University of Notre Dame, 390 Corbett Family Hall, Notre Dame, IN, 46556; email: ycheng4@nd.edu

Introduction

Differential item functioning (DIF) occurs when individuals of the same true latent ability or psychological trait from different demographic populations are found to have different chances of endorsing an item category (Holland & Wainer, 1993). For example, suppose a question on a national U.S. math assessment uses an idiom well known to individuals whose native language is English. Given the same math ability, individuals of English as a second language would have a lower chance of answering the item correctly due to their lack of familiarity with the idiom. Even after the test designer's best effort, problematic items such as this are unlikely to be identified by a simple visual check. Using statistical procedures to properly identify such items on an educational assessment helps secure fairness and impartiality of the assessment (Zwick, 2012).

The issue of DIF is distinct from inherent latent trait differences. High schoolers, for instance, will likely score higher on a math item than middle schoolers, which is a consequence of true ability differences on average (known as "impact"). A DIF effect occurs when differences are found after people have been matched on ability. If the item favors one subgroup at all ability levels, the effect is called uniform DIF. If the favored subgroup changes at different levels of the latent trait, the effect is known as non-uniform DIF. DIF assessment depends on the type of effect, and, in this study, we limit our discussion to the case of uniform DIF.

If an item truly functions differently between groups, the chance of identifying such item depends on many factors, including the sample size of each demographic group, average true latent trait scores in each group (impact), the DIF assessment method and the magnitude of the DIF effect (DeMars, 2011; Kubinger, Rasch, & Yanagida, 2009; Sireci & Rios, 2013; Zumbo, 1999). For DIF detection methods that depend on the selection of an *anchor set*, that is, a set of pre-selected test items that help establish a common metric between groups, the quality of the anchor is yet another crucial factor to successfully detect a DIF item. A threat to the quality of the anchor is *anchor contamination*, i.e., when an anchor contains one or more DIF items, which could hinder the ability to find a true DIF effect (Finch, 2005; Woods, 2009). If the anchor is contaminated, the common metric set by the anchor is inappropriate, and the latent trait estimates are biased (Holland & Thayer, 1988). Therefore, DIF research could greatly benefit from further understanding which factors affect the quality of the anchor. Previous studies have shown that two of these factors are anchor length and percentage of DIF items on the test. However, given a set of test items and its characteristics, i.e., difficulty and discrimination parameters, little is known about which items are most likely to be selected for the anchor, and how the anchor items' characteristics affect the power of DIF detection. In this study, our goal is to examine how item parameters influence the selection of anchor items, and subsequently, the DIF detection.

DIF assessment methods

A variety of statistical methods have been proposed to detect items with DIF within the test. Widely used DIF assessment techniques are the Mantel-Haenszel test (Holland & Thayer, 1988), the item-response-theory likelihood-ratio test (IRT-LRT; Thissen, Steinberg, & Wainer, 1988), the simultaneous item bias test (SIBTEST; Shealy & Stout, 1993) and the multiple indicators, multiple causes (MIMIC; Camilli & Shepard, 1994; Joreskog & Goldberger, 1975) model method. Unlike other methods, such as the Mantel-Haenszel method, which tests one item at a time based on contingency tables conditional on test sum-scores, the MIMIC model allows several items to be tested for DIF at once. Additionally, the structural equation modeling approach provides a flexible framework for testing both dichotomous and polytomous items simultaneously. Given these advantages, the MIMIC model approach has become increasingly popular for DIF detection (Chun, Stark, Kim, & Chernyshenko, 2016; Finch, 2005; Lee, Bulut, & Suh, 2017; Teresi, Ramirez, Lai, & Silver, 2008; Wang & Shih, 2010; Wang, Shih, & Yang, 2009; Woods, 2009; Woods & Grimm, 2011). We will, therefore, focus on applications of the MIMIC model in this paper.

In order to detect a DIF effect, a common metric must be established between the subgroups to allow group differences in the item to be compared against the common metric. Some methods that establish a common metric are: the equal-mean-difficulty method (Lord, 1980; Wang, 2004), which assumes the average item difficulty is the same for each subgroup; the all-other-items method (e.g., Cohen, Kim, & Wollack, 1996; Wang, 2004; Wang, Yeh, & Chia-Yi, 2003), which tests one item at a time while all other items on the test serve as matching variables (i.e., act as the anchor set); scale-purification methods (e.g., Hidalgo-Montesinos & Lopez-Pina, 2002; Lautenschlager, Flaherty, & Park, 1994; Wang et al., 2009), where items are included and removed from the matching set while testing the other items for DIF until the same set of items is flagged on two successive iterations; and the constant anchor item method (e.g., Wang, 2004), where a subset of items from the test that are most likely DIF-free is selected as the anchor based on their effect size or p-value (for a review of anchor methods, see Kopf, Zeileis, & Strobl, 2015).

DIF detection methods that rely on a previously chosen set of clean items, such as the constant anchor item method, have been called DIF-free-then-DIF methods (Wang, 2008; Wang, Shih, & Sun, 2012). They generally consist of two steps: in step 1, an anchor selection strategy is employed to select a set of DIF-free items to compose the anchor; in step 2, items outside the anchor are assessed for DIF. In this study, the DIF-free-then-DIF strategy is employed under the MIMIC model framework. In the first step, the anchor is selected through an iterative procedure with the MIMIC model (M-IT), and in the second step, DIF is assessed with a pure anchor (M-PA). These steps are described later in further detail.

Selecting an anchor set

In order to successfully identify DIF items in the second step of the DIF-free-then-DIF strategy, the anchor set selected in the first step must be clean, or free of DIF items. Among the factors known to affect anchor contamination are anchor length (number of anchor items) and percentage of DIF items on the test. As the number of DIF items on the test increases, the chance of anchor contamination also increases; thus, keeping the anchor short would reduce the possibility of erroneously including a DIF item in the anchor, especially when the test has many DIF items. Longer anchors yield generally higher power than a short one (Lopez Rivas, Stark, & Chernyshenko, 2009; Thissen et al., 1988). A four- or five-item anchor has been shown to reduce the chance of contamination while leading to desirable power rates when the test contains at most 40 % DIF items (Meade & Wright, 2012; Wang & Shih, 2010). Ideally, an anchor must be short enough to reduce anchor contamination but long enough to produce adequate power.

Although we understand what are desirable qualities in regards to anchor length and reducing chance of anchor contamination, much less is known about how item characteristics of the test affect anchor selection procedures, which may lead to higher chances of anchor contamination. For example, which items from a given test are more likely to be selected for the anchor? Moreover, how do items' characteristics affect anchor quality and further DIF detection? Lopez Rivas et al. (2009) studied the effects of anchor item discrimination and difficulty on IRT-LRT DIF detection and concluded that anchor quality and DIF assessment are both affected by item characteristics. A single-item anchor of high discrimination produced high power of DIF detection. For long anchors, the IRT-LRT was able to successfully identify DIF items as long as the anchor contained at least one highly discriminating item, even for conditions with a small DIF effect.

Similarly to Lopez Rivas et al. (2009), Meade and Wright (2012) constructed the anchor set by ranking items with non-significant DIF tests according to their discrimination parameter. They concluded that when using the IRT-LRT, an anchor made up of highly discriminating items yielded higher power rates than the all-other-items method. These two studies suggest that anchor item discrimination affects anchor quality and the power of DIF detection when using the IRT-LRT. Further investigation is needed about the role of anchor item difficulty. Anchor formulation will, of course, depend on the anchor selection strategy, and many strategies do not take the item parameters into account. In the MIMIC model framework, limited research has been done on what factors influence the formulation of the anchor, that is, which items are most likely to be included in the anchor.

Moreover, the current literature rarely addresses the relationship between DIF assessment and item parameters, with a few exceptions (e.g., Hong, 2010; Jodoin & Gierl, 2001; Sireci & Rios, 2013; Wang, 2008). For example, Jodoin and Gierl (2001) found inflated type-I error rates for hard and easy items when using the logistic regression procedure for DIF assessment; Hong (2010) concluded power of DIF detection decreases for items of small discrimination when using the MIMIC model.

Our main goal for the present study is to examine how item parameters of dichotomous items influence the selection of anchor items when the anchor is established through the M-IT, and subsequently, DIF detection through the M-PA. Both procedures are discussed in the next section.

Methods

The MIMIC model

Responses y_j to dichotomous items are conceptualized as realizations of an underlying continuous latent response variable y_j^*, such that:

$$y_j = \begin{cases} 1, & \text{if } y_j^* \geq \tau_j \\ 0, & \text{if } y_j^* < \tau_j \end{cases} \tag{1}$$

where $j = 1, ..., J$, with J the total number of items on the test, and τ_j the threshold for item j. τ_j represents the expected tendency of the item to be endorsed in the population. For example, considering an item with a standard normally distributed underlying continuous variable and $\tau_j = .7$, only about 25 % of the population has a tendency of answering the item correctly ($y_j = 1$).

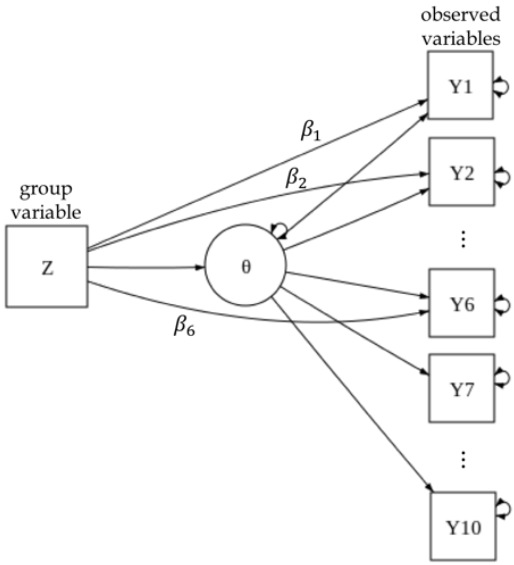


Figure 1:
The MIMIC model with 10 indicators, one "cause" variable and the group variable Z; DIF is tested on items 1 to 6 and the last four items Y_7, Y_8, Y_9 and Y_{10} act as the anchor set.

Under the structural equation modeling framework, the unidimensional MIMIC model is composed of the measurement:

$$\boldsymbol{y}^* = \boldsymbol{\Lambda}\theta + \boldsymbol{\beta}Z + \boldsymbol{\varepsilon} \tag{2}$$

and the structural model:

$$\theta = \gamma Z + \zeta \tag{3}$$

where $\boldsymbol{y}^* = (y_1^*, y_2^*, ..., y_J^*)'$ is the $J \times 1$ vector of underlying latent response variables; θ represents the person's latent trait variable; $\boldsymbol{\Lambda} = (\lambda_1, \lambda_2, ..., \lambda_J)'$ is the vector of factor loadings for each item; and $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_J)'$ is the vector of group path coefficients; Z is the group membership variable, where $Z = 1$ if the person belongs to the focal group and 0 otherwise; $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_J)'$ is the vector of measurement errors in $\boldsymbol{y}^*$. In eq. 3, γ represents the effect of group membership on the latent variable θ, i.e., it accounts for differences between groups in the average latent variable scores; ζ is the error variable.

If an item j is assumed to be free of DIF, β_j is set equal to zero for that item. Otherwise, DIF is tested by assessing whether β_j is significantly different from zero with the Wald test. Figure 1 illustrates DIF detection with the MIMIC model with ten indicators and one "cause" variable. Given one group variable, such as biological sex (male/female), and ten indicator variables, e.g., ten observed responses to questions on a math assessment, DIF is tested on all but four of the items, which have been previously selected as the anchor. Items 1 to 6 are tested for DIF, while items 7 to 10 act as the anchor set.

M-IT/M-PA procedure to assess DIF

The M-IT/M-PA is a DIF-free-then-DIF procedure which requires that an anchor set be selected first with the iterative MIMIC model (M-IT; Shih & Wang, 2009) before proceeding with DIF assessment through the MIMIC model with a pure anchor (M-PA; Wang & Shih, 2010). The M-IT establishes a common metric through the constant anchor item method, that is, it selects a constant number of items that are most likely DIF-free based on a rank of estimated β_j coefficients. An anchor set may contain, for instance, 1, 4 or 10 items; a four-item anchor has been shown to produce satisfactory results (Wang et al., 2012). Once a clean, DIF-free anchor is established, the MIMIC model with a short and pure anchor is employed to identify DIF items. The M-PA was found to be superior to other methods (equal-mean-difficulty and all-other-items) since it does not suffer from severe loss of power even when the test has as much as 40 % of DIF items. Below we provide technical details on the implementation of the DIF-free-then-DIF approach when using M-IT for anchor selection and M-PA for assessing DIF.

Anchor selection. In the first step, an anchor set is selected with the M-IT. Given a test with J items, the M-IT strategy of anchor selection takes the following steps:

1. Assume the first item is DIF-free (or "clean") and assess all other $J - 1$ items on the test for DIF;

2. Obtain one DIF index for each studied item (e.g., β_j), resulting in a total of $J-1$ DIF indices;
3. Repeat steps 1 and 2 assuming each item as DIF-free at a time, while testing all others for DIF; a total of $J-1$ DIF indices *per item* will be recorded;
4. Compute an average DIF index for each item and select the desired number of items with the smallest DIF indices to compose the anchor.

After the anchor has been established, items that are not part of the anchor are tested for DIF with the MIMIC model, which is the M-PA step.

DIF detection. In order to test items for DIF, the M-PA procedure is employed. Similarly to the model illustrated in Figure 1, given a J-item test and an anchor fixed at length 4, all J-4 items outside the anchor are tested simultaneously. While the anchor items have their group parameters β_j set to 0, the Wald test is used to assess whether the group parameters of the studied items are significantly different from 0 ($H_0 : \boldsymbol{\beta} = \mathbf{0}$). The Wald statistic is asymptotically χ^2-distributed with 1 degree of freedom under the null hypothesis (Dobson & Barnett, 2008), and the item is flagged if the test statistic is larger than the critical value.

When anchor methods are used, little is known about how item characteristics affect anchor formulation. Given a fixed DIF effect size, how will the item characteristics (more specifically item discrimination and difficulty parameters) affect anchor formulation and the subsequent DIF detection? To answer this question, we conduct a two-part simulation study that varies item parameters for the DIF and the DIF-free items and assesses items for DIF with the M-IT/M-PA, with an anchor of fixed length. In the first part of this study, we demonstrate the effect of item parameters on the selection of a clean anchor set, that is, an anchor that includes only DIF-free items, and show which items on the test are most likely to be selected for the anchor, given the item characteristics. In the second part of the study, assuming an anchor free of contamination, we assess how varying item parameters affect true- and false-positive rates of subsequent DIF detection. In summary, we investigate: (1) the relationship between item characteristics and anchor selection; and (2) the effects of item characteristics on DIF detection. This is done under the MIMIC model framework.

Simulation study: Part 1

The association between item characteristics and anchor selection is studied through a simulation study by performing the M-IT while varying the DIF item's discrimination and difficulty parameters.

Simulation design

Only one DIF item was included in the test so that the effect of the DIF item's discrimination and difficulty parameters on anchor selection may not be confounded with the

number of DIF items on the test. With 1 DIF item and 33 DIF-free item, the test length was fixed at 34 for every studied condition. Assuming the 2-parameter logistic model (2-PL; eq. 4 and 5), response data was generated for 500 examinees per group (focal/reference) – the total sample size was of 1000 examinees. True abilities of test-takers from the focal and the reference groups followed the standard normal distribution ($\theta \sim N(0,1)$).

For each iteration of the M-IT, each item acts as the interim matching variable at a time with their $\beta_j = 0$, while the MIMIC model is fit and a DIF index (absolute value of $\hat{\beta}_j$) is obtained for all other items on the test. With a total of 34 iterations, each item has 33 DIF indices, thus an average DIF index is computed for each item, $\overline{\hat{\beta}}_j = \Sigma_{t=1}^{33} \beta_j^{(t)}$. Items are then ranked on their average DIF indices, and the ones with the four smallest average indices are selected to compose the anchor. On the rare occasion of a tie, where the items ranked fourth and fifth (from smallest to largest) have the same average DIF index, the one with the smallest variance over iterations is selected for the anchor.

DIF effect

Under every condition, the DIF item is the first item of the test and has a uniform DIF effect favoring the reference group. As discussed previously, if a DIF effect is present, the item functions with different true values of the difficulty parameter for the focal and reference groups. Let the difficulty parameters of the focal and reference groups be b_{jF} and b_{jR}, respectively, for the DIF item. The DIF effect size in this study is chosen as the difference between difficulty parameters $\Delta b_j = b_{jF} - b_{jR}$. Under the 2-PL model, Raju (1988) showed that the signed area between the two ICCs is equivalent to the difference between the difficulty parameters in the two groups. There has been a long history of using the area between two item characteristic curves (ICCs) of reference versus focal groups as a measure of DIF effect, tracing at least back to the early 1980s (Rudner, Getson, & Knight, 1980; Shepard, Camilli, & Averill, 1981), and the difference between difficulty parameters has been used to generate DIF effect in many studies (e.g., Finch, 2005; Jamali, Ayatollahi, & Jafari, 2017; Jin, Myers, Ahn, & Penfield, 2012; Wang & Shih, 2010; Woods & Grimm, 2011). Therefore, we find it important to show its limitations in a case that applies to so many previous studies.

Given group membership and a fixed true ability level $\theta_R = \theta_F = \theta$, the probabilities of correctly answering an item when uniform DIF is present are:

$$P(y_{jR} = 1|\theta, R) = \frac{\exp(a_j(\theta - b_{jR}))}{1 + \exp(a_j(\theta - b_{jR}))}, \tag{4}$$

and

$$P(y_{jF} = 1|\theta, F) = \frac{\exp\left(a_j(\theta - (b_{jR} + \Delta b_j))\right)}{1 + \exp\left(a_j(\theta - (b_{jR} + \Delta b_j))\right)}, \tag{5}$$

where a_j is the discrimination parameter and y_{jR} and y_{jF} represent the observed responses for individuals of the reference and focal groups, respectively. For a DIF item with $\Delta b_j > 0$ ($\Delta b_j < 0$) the item will seem more difficult to individuals from the focal

(reference) group, given the same true ability level.

The item parameters from the 2-PL are related to the parameters of the MIMIC model through the equations (B. Muthén & Christoffersson, 1981):

$$a_j = \frac{\lambda_j}{\sqrt{1-\lambda_j^2}} \tag{6}$$

$$b_j = \frac{\tau_j - \beta_j Z}{\lambda_j} \tag{7}$$

when the latent variable is constrained to have mean 0 and variance 1. Additionally, if $Z = 0$ for the reference group respondents and $Z = 1$ for the focal group respondents, the effect size of uniform DIF may be rewritten as $\Delta b_j = b_{jF} - b_{jR} = -\beta_j/\lambda_j$ (MIMIC-ES; Jin et al., 2012). In what follows item parameters for the DIF item are indicated with the subscript "*DIF*". Similarly, item parameters of DIF-free, clean items are indicated with the subscript "*clean*".

A DIF effect was introduced by adding a difference of $\Delta b_{DIF} = .3, .5$ or .7 to the difficulty parameter of the focal group ($b_{F,DIF} = b_{R,DIF} + \Delta b_{DIF}$) for all simulation conditions. Given fixed a_{DIF} and Δb_{DIF}, the expected value of β_j can be derived through equations 6 and 7. Furthermore, the effects of .3, .5 and .7 are equivalent to negligible, moderate and large DIF effect sizes, respectively, on the delta scale (Zwick, 2012). Note that $\Delta_{MH} = -2.35 \times \Delta b_{DIF}$, therefore $\Delta_{MH} = -.705, -1.175$ and -1.645 for Δb_{DIF} values of .3, .5 and .7, respectively.

Item parameters

There was a total of 34 items on the test (1 DIF item and 33 DIF-free items). The DIF item was the first item on the test, and the remaining items were DIF free. Item parameters for the DIF-free items were fixed for every condition, while the DIF item parameters varied.

The DIF-free item parameters are described in Table 1. Item discrimination values were .5, 1.0 or 2.0, and the difficulty parameters were between -2.5 and 2.5 with increments of .5. The test had a total of eleven DIF-free items of each discrimination parameter value and three DIF-free items of the same difficulty parameter value. The DIF-free items parameters were varied this way to allow investigation of the number of times an item was selected through a range of item discrimination and difficulty values.

The DIF item parameter conditions are summarized in Table 2. Discrimination was .5, 1.0 or 2.0, and difficulty for the reference group varied between -2.5 and 2.0 with increments of .5. This fully-crossed design was chosen so that the effects of discrimination and difficulty parameters on DIF detection could be parsed apart. The DIF item parameters varied in a total of 3 discrimination $\times$ 10 difficulty $\times$ 3 DIF effect sizes = 90 conditions. Each simulation condition was replicated 1000 times.

Table 1:
Item difficulty and discrimination parameters for the 33 DIF-free (clean) items on the test.

Item	a_{clean}	b_{clean}	Item	a_{clean}	b_{clean}	Item	a_{clean}	b_{clean}
Y_2	.5	-2.5	Y_{13}	1.0	-2.5	Y_{24}	2.0	-2.5
Y_3	.5	-2.0	Y_{14}	1.0	-2.0	Y_{25}	2.0	-2.0
Y_4	.5	-1.5	Y_{15}	1.0	-1.5	Y_{26}	2.0	-1.5
Y_5	.5	-1.0	Y_{16}	1.0	-1.0	Y_{27}	2.0	-1.0
Y_6	.5	-.5	Y_{17}	1.0	-.5	Y_{28}	2.0	-.5
Y_7	.5	0	Y_{18}	1.0	0	Y_{29}	2.0	0
Y_8	.5	.5	Y_{19}	1.0	.5	Y_{30}	2.0	.5
Y_9	.5	1.0	Y_{20}	1.0	1.0	Y_{31}	2.0	1.0
Y_{10}	.5	1.5	Y_{21}	1.0	1.5	Y_{32}	2.0	1.5
Y_{11}	.5	2.0	Y_{22}	1.0	2.0	Y_{33}	2.0	2.0
Y_{12}	.5	2.5	Y_{23}	1.0	2.5	Y_{34}	2.0	2.5

Table 2:
DIF item (Y_1) difficulty and discrimination values: 30 simulation conditions.

Discrimination	Difficulty			
(a_{DIF})	$b_{R,DIF}$	$b_{F,DIF} = .3$	$b_{F,DIF} = .5$	$b_{F,DIF} = .7$
.5	-2.5	-2.2	-2.0	-1.8
1.0	-2.0	-1.7	-1.5	-1.3
2.0	-1.5	-1.2	-1.0	-.8
	-1.0	-.7	-.5	-.3
	-.5	-.2	0	.2
	0	.3	.5	.7
	.5	.8	1.0	1.2
	1.0	1.3	1.5	1.7
	1.5	1.8	2.0	2.2
	2.0	2.3	2.5	2.7

Outcome variables

Anchor frequency. In the M-IT step, four items are selected for the anchor set, and for each replication, we record which items are chosen. Anchor frequency is reported for each DIF-free item as the average number of times that item was selected into the anchor per DIF item discrimination (a_{DIF}) condition. If all items had the same characteristics, each item would be expected to be selected into the anchor $4/J \times 1000$ times, where J is the total number of items on the test. With $J = 34$, DIF-free items are each expected to be selected about 117 times out of 1000 replications. Anchor frequency is only reported for DIF-free items. Anchor accuracy rates provide information about whether the DIF

item ended up in the anchor.

Anchor accuracy. For each replication, we record whether the anchor set included the DIF item. Anchor accuracy is the proportion of times the anchor is clean; that is, it contains only DIF-free items.

The simulation study was conducted using R and Mplus. Data was generated with R version 3.3.2 (R Core Team, 2018). All subsequent analyses, including M-IT and M-PA, were run on Mplus (Muthén & Muthén, 1998-2017) through the *MplusAutomation* package in R (Hallquist & Wiley, 2018). The default weighted least squares estimator in Mplus (`ESTIMATOR=WLSMV`) was used (B. O. Muthén, du Toit, & Spisic, 1997).

Results

Anchor frequency. Tables 3 and 4 show average anchor frequencies (across DIF item difficulty conditions) for DIF-free items averaged over their discrimination and difficulty parameters conditions, respectively. In these two tables, the DIF item had a moderate effect size of $\Delta b_{DIF} = .5$. Table 3 shows that clean items with small discrimination parameters were more likely to be selected into the anchor. For example, when the DIF item had a discrimination parameter $a_{DIF} = 1.0$, the average anchor frequencies were 362, 2 and 0 when the clean items were low, moderate or highly discriminating, respectively. Discrimination parameter of the DIF item had no effect on anchor frequencies. Table 4 shows that average anchor frequencies for clean items remained about the same across difficulty parameter values. For instance, when the DIF item discrimination parameter was $a_{DIF} = 1.0$, average frequencies were 122 and 121 for DIF-free items of $b_{clean} = 0$ and $b_{clean} = 2.0$, respectively. In addition, the discrimination parameter of the DIF item does not seem to affect anchor selection frequency. Results for the conditions of negligible ($\Delta b_{DIF} = .3$) and large ($\Delta b_{DIF} = .7$) DIF effect sizes are reported in the Appendix in Tables 7 to 10 and similar trends were found under those conditions. DIF magnitude did not seem to affect anchor selection frequency.

Table 3:
Average frequency of DIF-free items selected by the M-IT per item discrimination parameter ($\Delta b_{DIF} = .5$).

	a_{clean}		
a_{DIF}	.5	1.0	2.0
.5	354	2	0
1.0	362	2	0
2.0	362	2	0

Table 4:
Average frequency of DIF-free items selected by the M-IT per item difficulty parameter ($\Delta b_{DIF} = .5$).

	b_{clean}										
a_{DIF}	-2.5	-2.0	-1.5	-1.0	$-.5$	0	.5	1.0	1.5	2.0	2.5
.5	119	121	121	119	116	117	117	120	116	122	119
1.0	122	121	121	121	121	122	119	119	122	121	125
2.0	123	120	122	120	120	120	120	122	120	123	124

Anchor accuracy. The DIF item was sometimes included by the M-IT in the anchor set. Rates of anchor accuracy are reported in Table 5. Accuracy rates were 100 % under all conditions as long as the item was moderate ($a_{DIF} = 1.0$) to highly discriminating ($a_{DIF} = 2.0$). When the DIF item had small discrimination $a_{DIF} = .5$, rates of accuracy ranged from 89 to 94 % and were smaller for items with extreme difficulty parameter values, that is, when the item was very easy or very hard. For example, when $b_{R,DIF} = -2.0$, the anchor set was clean 91 % of the time, against 94 % accuracy when $b_{R,DIF} = 0$. In summary, large discrimination and moderate difficulty of the DIF item were associated with higher accuracy rates.

Simulation study: Part 2

In the second part of this study, we evaluate the performance of the M-PA with a pure four-item anchor in terms of true- and false-positive rates of DIF detection with three anchor configurations (small, medium and large anchor item discrimination) and varied DIF item parameters.

Simulation design

Test length, sample size, item parameters and DIF effects remain the same as in Part 1. Anchor sets were constructed from a subset of four items from the test. Out of a

Table 5:
Accuracy rates of selecting a clean anchor with the M-IT ($\Delta b_{DIF} = .5$).

	Difficulty									
$b_{R,DIF}$	-2.5	-2.0	-1.5	-1.0	$-.5$	0	.5	1.0	1.5	2.0
$b_{F,DIF}$	-2.0	-1.5	-1.0	$-.5$	0	.5	1.0	1.5	2.0	2.5
a_{DIF}	Accuracy Rates									
.5	.893	.906	.910	.924	.940	.940	.938	.924	.907	.918
1.0	1	1	1	1	1	1	1	1	1	1
2.0	1	1	1	1	1	1	1	1	1	1

test with 33 clean items, four clean items were chosen as the anchor according to three configurations: *small* ($a_{anchor} = .5$), *medium* ($a_{anchor} = 1.0$) and *large* ($a_{anchor} = 2.0$) discrimination parameters. For each item discrimination condition, the anchor items were the ones with difficulty parameters -1.0, -.5, .5 and 1.0. That is, as displayed in Table 1, the items that composed the small anchor item discrimination condition were Y_5, Y_6, Y_8 and Y_9; for the medium discrimination condition, items Y_{16}, Y_{17}, Y_{19} and Y_{20}; and, for the large discrimination condition, items Y_{27}, Y_{28}, Y_{30} and Y_{31}. We only examined the effect of varying anchor items' discrimination parameter on power of DIF detection, as we expect the effect of anchor item discrimination to be more pronounced than anchor item difficulty.

Outcome variables

Type-I error. Type-I error rates are computed as the proportion of times the DIF test of a clean item is significant out of 1000 replications for each of the 29 clean items outside the anchor. Type-I error rates were computed after running the M-PA for each simulation condition, and therefore, type-I error rates are averaged over all conditions of DIF item parameters and DIF effect sizes, as no effect on the rates of false positives is expected from varying those conditions.

Power. The proportion of significant tests of the DIF item out of 1000 replications is reported across item parameters and anchor configurations conditions.

Results

Type-I error rates. Average type-I error rates are reported in Table 6 for each clean item discrimination parameter value for an anchor of small ($a_{anchor} = .5$), medium ($a_{anchor} = 1.0$) and large ($a_{anchor} = 2.0$) item discrimination parameter. Each number in Table 6 represents an average over items with the same discrimination value. For example, the first element of the first row is an average rate over DIF tests of items Y_2 to Y_4, Y_7, and Y_{10} to Y_{12}, given a small anchor item discrimination. The diagonal cells are type-I error rates averaged over items that have the same discrimination as the anchor items, and therefore represent an average over seven items. All off-diagonal elements are

averages over eleven items within the test that have the same discrimination parameter (none of which were part of the anchor). Results show that, in general, false positive rates are well-controlled for each anchor configuration condition.

Table 6:
Average type-I error (and standard deviation) for each clean item discrimination parameter condition given anchor configuration.

		DIF-free item discrimination		
Anchor configuration	a_{anchor}	$a_{clean} = .5$	$a_{clean} = 1.0$	$a_{clean} = 2.0$
small	.5	$.048^{\dagger}(.006)$	$.048(.006)$	$.049(.007)$
medium	1.0	$.047(.007)$	$.049^{\dagger}(.007)$	$.051(.007)$
large	2.0	$.050(.007)$	$.049(.007)$	$.049^{\dagger}(.006)$

†Type-I error rates are averaged over seven clean items that are not in the anchor but have the same a as the anchor items. All other values in the table are averaged over a total of eleven items.

Power rates. Due to space limitations, we report power rates here only for a moderate DIF effect ($\Delta b_{DIF} = .5$) and a medium anchor item discrimination ($a_{anchor} = 1.0$), which are shown in Figure 2. A constant relationship is found between DIF item difficulty and power when $a_{DIF} = .5$ and a quadratic relationship when $a_{DIF} = 1.0$ or 2.0. The proportion of times the DIF item was correctly identified was overall higher when the DIF item had a large discrimination parameter. For example, power rates were .45, .88 and .99 for $a_{DIF} = .5$, $a_{DIF} = 1.0$ and $a_{DIF} = 2.0$, respectively, when $b_{DIF} = 0$. When the item was too easy/difficult, power rates were smaller than when the items had mid-range difficulty, unless the DIF item discrimination was very low ($a_{DIF} = .5$). Complete results can be found in the Appendix for the small (Figures 4 to 6), medium (Figures 7 and 8) and large (Figures 9 to 11) anchor item discrimination conditions. Overall, the same trends for power of DIF detection are observed for other conditions of anchor item discrimination and DIF effect sizes. As expected, the DIF item is more easily identified by the M-PA as DIF effect size increases, given a fixed discrimination and difficulty parameter value.

Anchor configuration also affected the power of DIF detection, as shown by Figure 3, where results for a DIF item of moderate DIF effect ($\Delta b_{DIF} = .5$) are reported for three levels of item difficulty of the reference group: a) medium ($b_{R,DIF} = 0$), b) medium-high ($b_{R,DIF} = 1.0$) and c) high ($b_{R,DIF} = 2.0$). In general, power decreased as the DIF item difficulty level increased, while the power increased with increasing values of DIF item discrimination. For all levels of DIF item difficulty, given a fixed DIF item discrimination, the power rates were larger if the anchor items were highly discriminating. For example, in Figure 3a, when $a_{DIF} = 1.0$, the empirical power rates were .73, .88 and .93 for the anchor item small-, medium- and large-discrimination conditions, respectively. Due to space limitations, we omit the results for other studied conditions of $b_{R,DIF}$. Note that if $b_{R,DIF} = -1.0$ or -2.0, power rates are expected

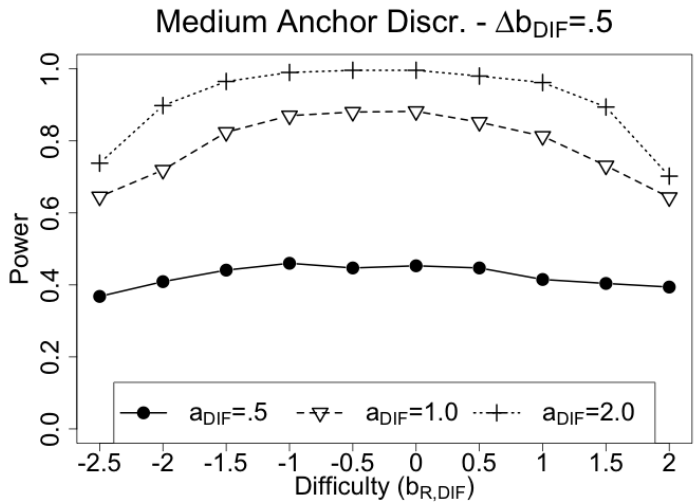


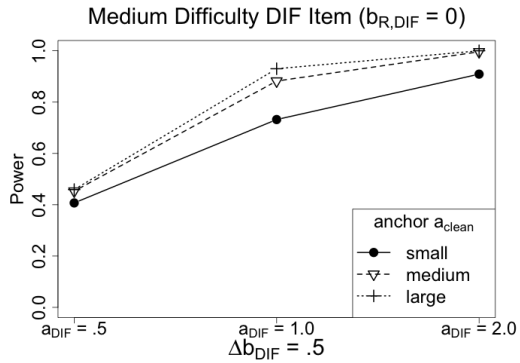
Figure 2:
Power rates given a clean anchor set with items of medium discrimination ($a_{anchor} = 1.0$) and moderate DIF effect size ($\Delta b_{DIF} = .5$).

to be similar to those shown for $b_{R,DIF}$ of equal absolute value but different direction. Results for the conditions where Δb_{DIF} was .3 or .7 are reported in the Appendix in Figures 12 and 13 and similar trends were observed for each of the DIF effect size conditions.

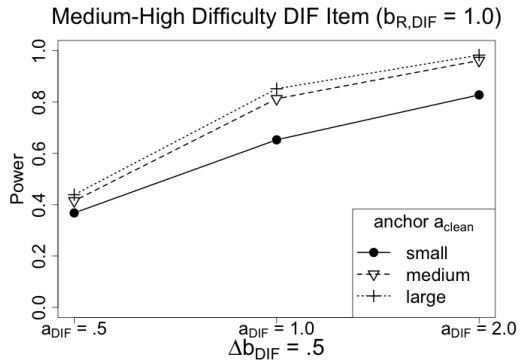
Conclusions

Although anchor contamination and the quality of the anchor set may affect our ability to identify DIF items, few studies have considered the effect of difficulty and discrimination parameters on anchor selection and subsequent DIF detection. Through a simulation study, we show how item characteristics affect (1) anchor selection and (2) true- and false-positive rates of DIF detection.

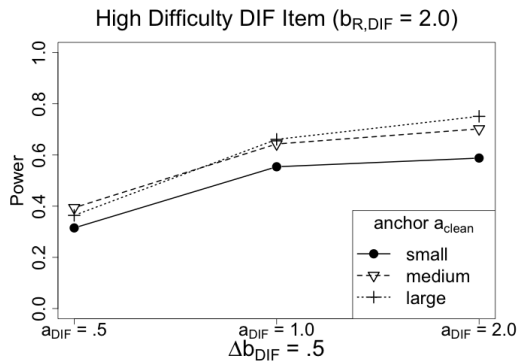
Results from the simulation study show that the anchor set produced by the iterative MIMIC model procedure (M-IT) is more likely to include only DIF-free items if the DIF item has a large discrimination parameter. That is, the M-IT will rarely select DIF items with high discrimination for the anchor, and the risk of anchor contamination is therefore decreased. These results are not surprising if we consider the mechanism underlying the anchor selection procedure and how the DIF indices change with the item parameters. The M-IT is a rank-based method that ranks items according to their estimated average DIF index (absolute group parameter $\hat{\beta}$) and selects the items with the m smallest DIF indices ($m = 4$ in this study). A large discrimination parameter



(a)



(b)



(c)

Figure 3: Power rates across anchor configuration conditions for a moderate DIF effect size ($\Delta b_{DIF} = .5$) when (a) $b_{R,DIF} = 0$, (b) $b_{R,DIF} = 1.0$ and (c) $b_{R,DIF} = 2.0$.

amplifies the DIF index, that is, given two items with the same DIF effect $b_F - b_R$, the item with a large discrimination parameter is expected to have greater DIF indices ($|\hat{\beta}|$) than the one with a small discrimination parameter. This can be shown analytically for the MIMIC model. Given the effect size MIMIC-ES for uniform DIF, $\Delta b = \frac{-\beta}{\lambda}$, some algebraic manipulation and equation (6) lead to the following relationship between β and a:

$$\beta = -\Delta b\, \lambda = -\Delta b \frac{a}{\sqrt{1+a^2}}. \tag{8}$$

Therefore, β estimates will increase (in absolute value) as a increases, even if the DIF effect size Δb remains the same. In this study, we considered the case of only one DIF item on the test; thus, across conditions, while keeping everything else constant, a highly discriminating DIF item is less likely to be selected for the anchor than a DIF item of small discrimination. Conversely, a DIF item with low discrimination is more likely to be mistakenly included in the anchor set.

A similar trend is found for the DIF-free items most likely to be selected into the anchor. If a clean item has a small discrimination parameter, it is more frequently chosen for the anchor than a clean item of high discrimination. Intuitively, this result can be explained by the observed discrepancies between groups resulting from sampling variability. Such discrepancies make it seem as if an item of high discrimination functions differently between groups, more so than an item of low discrimination, at each ability level, making it more likely for the highly discriminating item to be ranked lower in the M-IT procedure. Sampling variability also makes it more likely for a DIF-free item with large discrimination parameter to appear to have a DIF effect than truly DIF items of low discrimination, and consequently, for DIF items to be chosen over DIF-free items for the anchor. No effect was found regarding item difficulty on the frequency which clean items are selected for the anchor. Clean items of varying levels of difficulty were selected for the anchor with about the same frequency.

Accuracy rates were 100 % across all conditions as long as the DIF item had at least a moderate effect ($\Delta b_{DIF} = .5$) and a medium ($a_{DIF} = 1.0$) or large ($a_{DIF} = 2.0$) discrimination parameter. For a DIF item of low discrimination ($a_{DIF} = .5$), the chance of anchor contamination increases as the item difficulty level goes to the extremes. This shows a small but relevant effect of item difficulty on the quality of the anchor.

In summary, part 1 of this study suggests that item discrimination plays an important role in anchor selection. Part 1 showed that clean items of low discrimination are more likely to be selected for the anchor than highly discriminating ones. Therefore, in part 2, we address the question as to how anchor item discrimination affects the power of DIF detection. Overall, our results suggest that a small anchor item discrimination negatively impacts power of DIF detection; with the assumption of a clean anchor, the DIF item is more easily identified if the items composing the anchor are highly discriminating.

Furthermore, the power of DIF detection also depends on the DIF item difficulty parameter, but such effect changes at different levels of item discrimination. For DIF items

with very small discrimination parameters (.5), power remains about the same for any value of difficulty of the DIF item, given fixed DIF magnitude and anchor configuration. When the DIF item has a moderate to large discrimination parameter (1.0 or 2.0), power decreased quadratically with increasing difficulty parameters. That is, items had higher power for the mid-range difficulty levels, and the lowest power was found when the items were very easy or very difficult. Because difficulty and discrimination parameters are usually positively correlated, we would rarely encounter an item that is very easy (large negative value) *and* highly discriminating. Therefore, we are mainly interested in the cases where, for example, easy items have small discrimination or hard items have large discrimination. In such cases, for the purpose of DIF assessment, we would like to avoid items that are too easy, since they would likely not discriminate well between different ability levels, which would make DIF testing a challenge, while items in the range of medium to high difficulty would provide acceptable levels of discrimination and consequently of power of DIF detection.

Anchor configuration also affected the power of DIF detection. An anchor composed of highly discriminating items produced the largest power for nearly all conditions. The M-IT, however, more frequently selects items of low discrimination, which could hinder the ability to identify DIF items.

Discussion

A possible solution to the issue of frequent selection of items of low discrimination would be to implement a similar anchor selection strategy like the one employed by Lopez Rivas et al. (2009) and Meade and Wright (2012), where the anchor was constructed by ranking items with non-significant DIF tests according to their discrimination parameter. An application of such a strategy to the MIMIC model method of assessing DIF would produce anchors with items of high discrimination while taking advantage of the flexibilities of the MIMIC model framework.

In simulation studies, the DIF effect size (for example, Raju's signed area $b_F - b_R$) is usually fixed across DIF items on the same test, and the *average* power rates of successfully identifying DIF items are reported (e.g., Wang et al., 2009). This requires interpreting power values with caution because they depend on the particular item parameters in the assessment. When multiple DIF items are considered in a simulation study, an average power rate may be reported, and, in that case, the items' parameters are usually not equal across items. For example, Shih and Wang (2009) and Wang et al. (2009) had as many as 15 DIF items on a 50-item test, in which case discrimination parameters ranged from .7 to 2.0 and had an average value of 1.27 (both studies used the item parameters derived in Cohen et al., 1996). Woods and Grimm (2011) reported average power rates for uniform DIF items, which comprised of a third of tests with varying lengths of 6, 12 or 24. In that study, discrimination parameters were randomly generated from $N(\mu = 1.7, \sigma = .3)$ and truncated between .5 and 2.0. Although a test with different values for the discrimination parameter conforms to the expectation that

items within the same test may not uniformly discriminate the population, reporting the average power of DIF detection can be misleading. Simulation studies that report average power but do not stress the difference in discrimination parameter values may lead readers to expect more/less power from a DIF method than it is warranted, given the practitioner's own test item characteristics. Additionally, a DIF researcher may not be able to compare DIF studies on the same method. For clarity in interpretation, we recommend that future studies report power conditional on the item parameters or explicitly report on the DIF item parameter values and their relationship with power of DIF detection.

This study has some limitations. All simulations focused on anchor selection and uniform DIF detection of dichotomous items when using the M-IT/M-PA two-step procedure. Other anchor selection procedures and other DIF tests (e.g., Mantel-Haenszel test or SIBTEST) for uniform and non-uniform DIF on dichotomous or polytomous items may be studied in the future. Additionally, the measure of DIF effect size chosen for this study (Raju's signed area) has three main limitations: first, its application to non-uniform DIF may be problematic in the case where the item characteristic curves cross near the difficulty parameter, as the areas would cancel each other out; second, it does not have a one-to-one relationship to the power of identifying a DIF effect; and third, it does not measure DIF effect size accurately in the presence of impact (DeMars, 2011). This study investigates tests with a sample size of $N = 1000$ and assumes focal and reference groups are of equal size, but in practice, the focal group is expected to be smaller than the reference group. Furthermore, it is generally expected for a test to have several DIF items, while this study only investigated the condition where 1 out of 34 items on the assessment had a DIF effect. This study design was chosen out of necessity, as including more DIF items would make interpreting the results dependent on the *average* DIF items' discrimination and difficulty parameters. As the number of DIF items increases, interpretation could become increasingly challenging. Future directions include evaluating other anchor selection and DIF methods, while varying sample sizes, reference and focal group sample size ratio, and mean ability difference between the two groups.

Finally, this study points to a gap in the current DIF literature: a lack of a widely used and accepted effect size measure for DIF effect. DeMars (2011) suggests that power is associated with the measure of DIF magnitude. The author discusses the relationship between item parameters and existing measures of DIF magnitude and theoretically demonstrates that different measures of DIF effect are based on different metrics and, therefore, relate to the item parameters in a variety of ways. For instance, the author showed that the log-odds ratio increases as the item discrimination increases but remains constant for varying item difficulty values. Thus, the Mantel-Haenszel-based effect size is larger for highly discriminating items than for poorly discriminating items. In turn, the IRT-based probability of correct difference between reference and focal group decreases for items of low discrimination value and items of extreme difficulty parameter values, given the same DIF magnitude. This also has implications for the standardized

p-difference (Dorans, Schmitt, & Bleistein, 1988), which is based on the difference in probabilities of correct for each demographic group weighted by the number of respondents in each ability level (e.g., sum-score levels) for the focal group. Although DeMars (2011) did not directly evaluate power of DIF detection, we expect that DIF items with large magnitude will be more easily identified by the DIF method than items of small magnitude. In other words, power of DIF detection will depend on the DIF effect size measure and the scale set by such measure. Determining a standard, widely-used DIF effect size measure would allow researchers to more clearly interpret and compare DIF studies results, regardless of the technique used, and is a potential future direction from this study.

References

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016, oct). MIMIC methods for detecting DIF among multiple groups. *Applied Psychological Measurement*, *40*(7), 486–499. doi: 10.1177/0146621616659738

Cohen, A. S., Kim, S.-H., & Wollack, J. A. (1996, mar). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement*, *20*(1), 15–26. doi: 10.1177/014662169602000102

DeMars, C. E. (2011). An analytic comparison of effect sizes for differential item functioning. *Applied Measurement in Education*, *24*(3), 189–209. doi: 10.1080/08957347.2011.580255

Dobson, A. J., & Barnett, A. (2008). *An Introduction to Generalized Linear Models* (3rd ed.). doi: 10.2307/2348299

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1988). *The standardization approach to assessing differential speedness* (Tech. Rep. No. May). Princeton, New Jersey: Educational Testing Service.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278–295. doi: 10.1177/0146621605275728

Hallquist, M. N., & Wiley, J. F. (2018, jul). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638. doi: 10.1080/10705511.2017.1402334

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*(1), 32–44. doi: 10.1177/0013164402062001003

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, US: Lawrence Erlbaum Associates, Inc.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Hong, T. (2010). *The utility of the MIMIC model and MCFA method when detecting DIF using Monte Carlo simulation* (Dissertation). Purdue University.

Jamali, J., Ayatollahi, S. M. T., & Jafari, P. (2017). The effect of small sample size on measurement equivalence of psychometric questionnaires in MIMIC model: a simulation study. *BioMed Research International*, *2017*, 1–12. doi: 10.1155/2017/7596101

Jin, Y., Myers, N. D., Ahn, S., & Penfield, R. D. (2012). A comparison of uniform DIF effect size estimators under the MIMIC and Rasch models. *Educational and Psychological Measurement*, *73*(2), 339–358. doi: 10.1177/0013164412462705

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type-I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349. doi: 10.1207/S15324818AME1404_2

Joreskog, K. G., & Goldberger, A. S. (1975, sep). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351), 631–639. doi: 10.2307/2285946

Kopf, J., Zeileis, A., & Strobl, C. (2015). A framework for anchor methods and an iterative forward approach for DIF detection. *Applied Psychological Measurement*, *39*(2), 83–103. doi: 10.1177/0146621614544195

Kubinger, K. D., Rasch, D., & Yanagida, T. (2009). On designing data-sampling for Rasch model calibrating an achievement test. *Psychology Science Quaterly*, *51*(4), 370-384.

Lautenschlager, G. J., Flaherty, V. L., & Park, D. G. (1994). IRT differential item functioning: An examination of ability scale purifications. *Educational and Psychological Measurement*, *54*(1), 21–31. doi: 10.1177/0013164494054001003

Lee, S., Bulut, O., & Suh, Y. (2017, aug). Multidimensional extension of multiple indicators multiple causes models to detect DIF. *Educational and Psychological Measurement*, *77*(4), 545–569. doi: 10.1177/0013164416651116

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement*, *33*(4), 251–265. doi: 10.1177/0146621608321760

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

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*(5), 1016–1031. doi: 10.1037/a0027934

Muthén, B., & Christoffersson, A. (1981, dec). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, *46*(4), 407–419. doi: 10.1007/BF02293798

Muthén, B. O., du Toit, H. C. S., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes* (Tech. Rep.). Mplus.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from `https://www.r-project.org/`

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics*, *5*(3), 213-233. doi: 10.2307/1164965

Shealy, R., & Stout, W. (1993, jun). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194. doi: 10.1007/BF02294572

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, *6*(4), 317-375.

Shih, C.-L., & Wang, W.-C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, *33*(3), 184–199. doi: 10.1177/0146621608321758

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2-3), 170–187. doi: 10.1080/13803611.2013.767621

Teresi, J. A., Ramirez, M., Lai, J.-S., & Silver, S. (2008). Occurrences and sources of differential item functioning (DIF) in patient-reported outcome measures: description of DIF methods, and review of measures of depression, quality of life and general health. *Psychology Science Quarterly*, *50*(4), 538.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education*, *72*(3), 221–261. doi: 10.3200/JEXE.72.3.221-261

Wang, W.-C. (2008). Assessment of differential item functioning. *Journal of Applied Measurement*, *9*(4), 387–408.

Wang, W.-C., & Shih, C.-L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement*, *34*(3), 166–180. doi: 10.1177/0146621609355279

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*(4), 687–708. doi: 10.1177/0013164411426157

Wang, W.-C., Shih, C.-L., & Yang, C.-C. (2009, oct). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement*, *69*(5), 713–731. doi: 10.1177/0013164409332228

Wang, W.-C., Yeh, Y.-L., & Chia-Yi. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479–498. doi: 10.1177/0146621603259902

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57. doi: 10.1177/0146621607314044

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*(5), 339–361. doi: 10.1177/0146621611405984

Zumbo, B. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*(January 1999), 1–57.

Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: flagging rules, minimum sample size requirements, and criterion refinement. *ETS Research Report Series*, *2012*(1). doi: 10.1002/j.2333-8504.2012.tb02290.x

Appendix

Table 7:
Average frequency of DIF-free items selected by the M-IT per item discrimination parameter ($\Delta b_{DIF} = .3$).

	a_{clean}		
a_{DIF}	.5	1.0	2.0
.5	344	2	0
1.0	361	3	0
2.0	361	3	0

Table 8:
Average frequency of DIF-free items selected by the M-IT per item discrimination parameter ($\Delta b_{DIF} = .7$).

	a_{clean}		
a_{DIF}	.5	1.0	2.0
.5	360	2	0
1.0	362	2	0
2.0	362	2	0

Table 9:
Average frequency of DIF-free items selected by the M-IT per item difficulty parameter ($\Delta b_{DIF} = .3$).

	b_{clean}										
a_{DIF}	-2.5	-2.0	-1.5	-1.0	$-.5$	0	.5	1.0	1.5	2.0	2.5
.5	119	115	117	114	114	114	112	115	114	117	117
1.0	123	121	121	121	120	122	118	119	119	125	124
2.0	126	122	121	118	122	119	118	121	118	121	125

Table 10:
Average frequency of DIF-free items selected by the M-IT per item difficulty parameter ($\Delta b_{DIF} = .7$).

	b_{clean}										
a_{DIF}	-2.5	-2.0	-1.5	-1.0	$-.5$	0	.5	1.0	1.5	2.0	2.5
.5	128	119	121	121	118	121	118	117	121	120	123
1.0	123	122	124	120	121	119	118	117	122	123	126
2.0	123	122	120	121	116	122	120	121	119	124	125

Table 11:
Accuracy rates of selecting a clean anchor with the M-IT ($\Delta b_{DIF} = .3$).

	Difficulty									
$b_{R,DIF}$	-2.5	-2.0	-1.5	-1.0	$-.5$	0	.5	1.0	1.5	2.0
$b_{F,DIF}$	-2.2	-1.7	-1.2	$-.7$	$-.2$	.3	.8	1.3	1.8	2.3
a_{DIF}	Accuracy Rates									
.5	.788	.809	.782	.797	.835	.807	.813	.809	.804	.815
1.0	.998	1	1	1	.999	1	1	1	1	.999
2.0	1	1	1	1	1	1	1	1	1	1

Table 12:
Accuracy rates of selecting a clean anchor with the M-IT ($\Delta b_{DIF} = .7$).

	Difficulty									
$b_{R,DIF}$	-2.5	-2.0	-1.5	-1.0	$-.5$	0	.5	1.0	1.5	2.0
$b_{F,DIF}$	-1.8	-1.3	$-.8$	$-.3$	.2	.7	1.2	1.7	2.2	2.7
a_{DIF}	Accuracy Rates									
.5	.972	.983	.987	.984	.990	.985	.978	.987	.985	.964
1.0	1	1	1	1	1	1	1	1	1	1
2.0	1	1	1	1	1	1	1	1	1	1

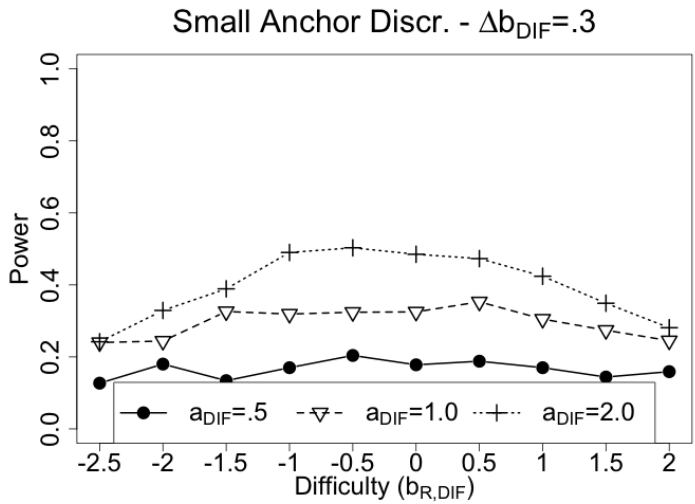


Figure 4:
Power rates for a 34-item test, given a clean anchor set with items of small discrimination ($a_{anchor} = .5$) and negligible DIF effect size ($\Delta b_{DIF} = .3$).

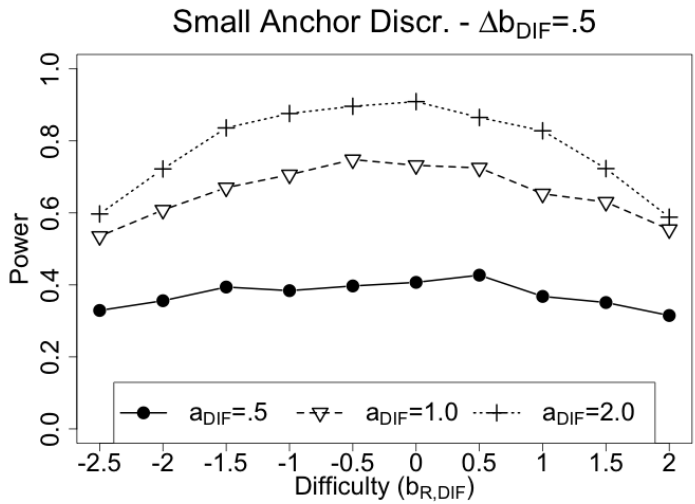


Figure 5:
Power rates for a 34-item test, given a clean anchor set with items of small discrimination ($a_{anchor} = .5$) and moderate DIF effect size ($\Delta b_{DIF} = .5$).

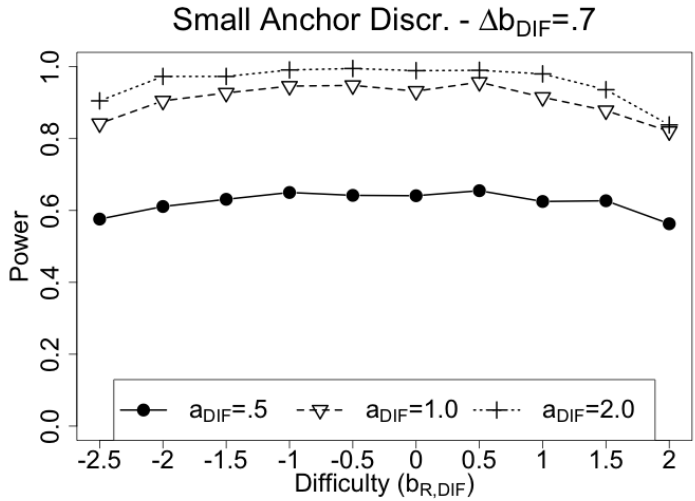


Figure 6:
Power rates for a 34-item test, given a clean anchor set with items of small discrimination ($a_{anchor} = .5$) and large DIF effect size ($\Delta b_{DIF} = .7$).

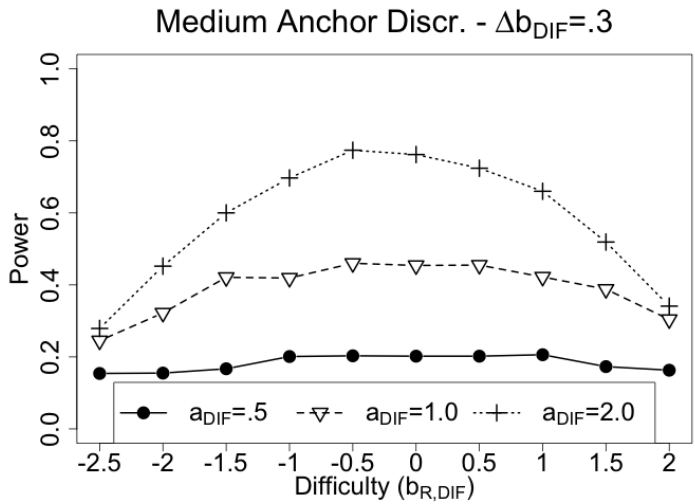


Figure 7:
Power rates for a 34-item test, given a clean anchor set with items of medium discrimination ($a_{anchor} = 1.0$) and negligible DIF effect size ($\Delta b_{DIF} = .3$).

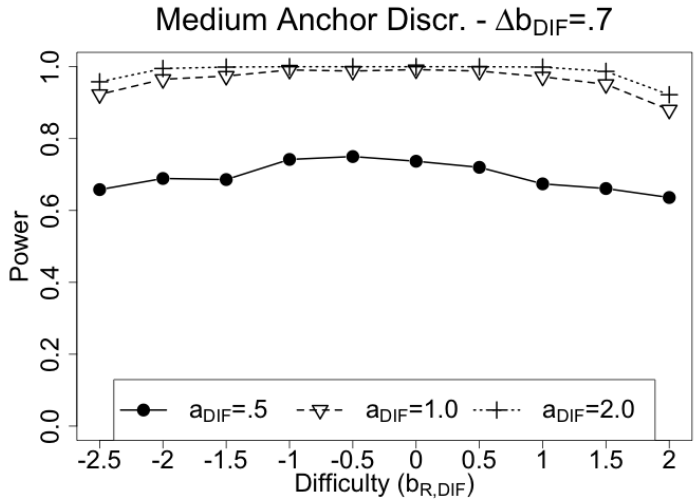


Figure 8:
Power rates for a 34-item test, given a clean anchor set with items of medium discrimination ($a_{anchor} = 2.0$) and large DIF effect size ($\Delta b_{DIF} = .7$).

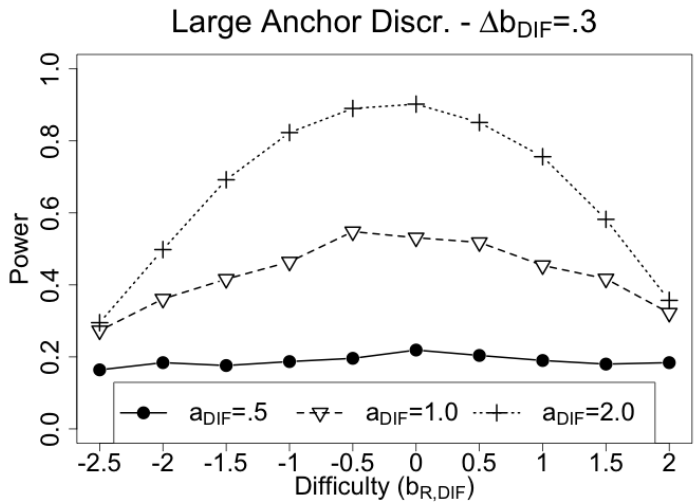


Figure 9:
Power rates for a 34-item test, given a clean anchor set with items of large discrimination ($a_{anchor} = 2.0$) and negligible DIF effect size ($\Delta b_{DIF} = .3$).

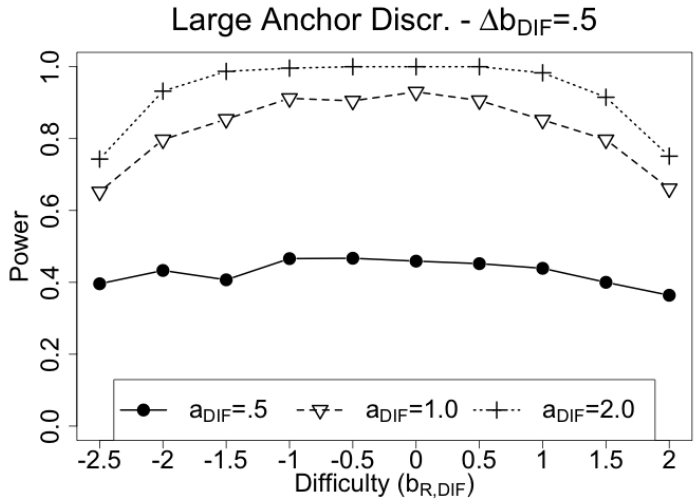


Figure 10:
Power rates for a 34-item test, given a clean anchor set with items of large discrimination ($a_{anchor} = 2.0$) and moderate DIF effect size ($\Delta b_{DIF} = .5$).

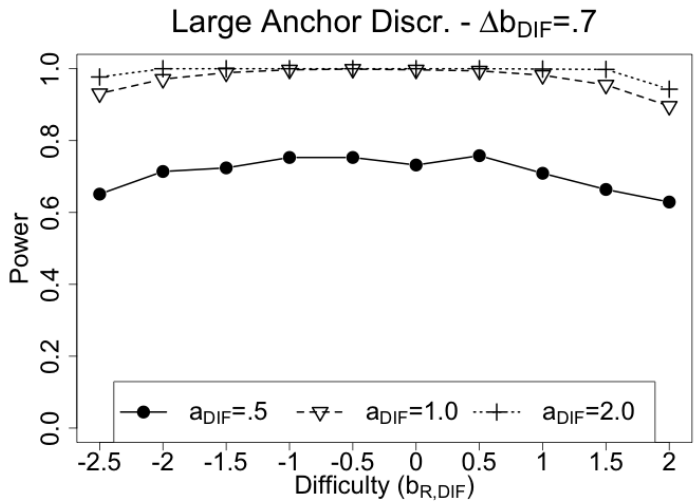
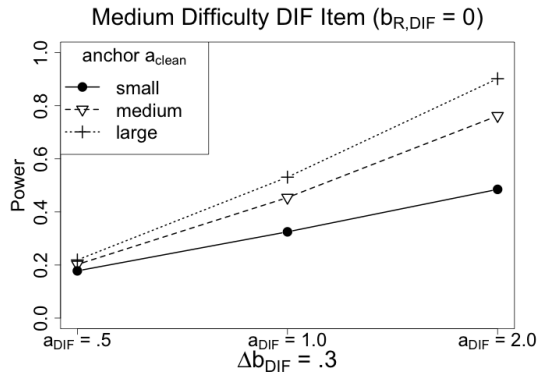
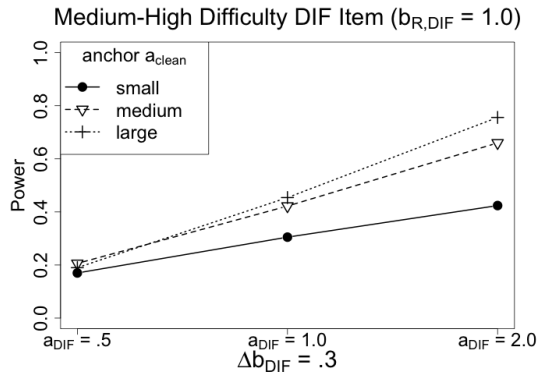


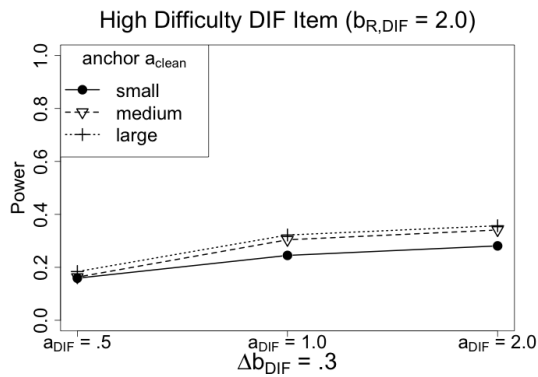
Figure 11:
Power rates for a 34-item test, given a clean anchor set with items of large discrimination ($a_{anchor} = 2.0$) and large DIF effect size.



(a)

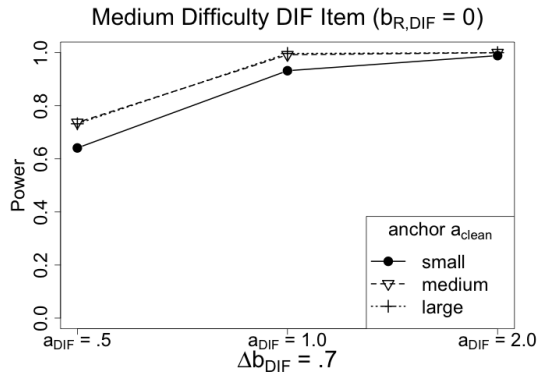


(b)

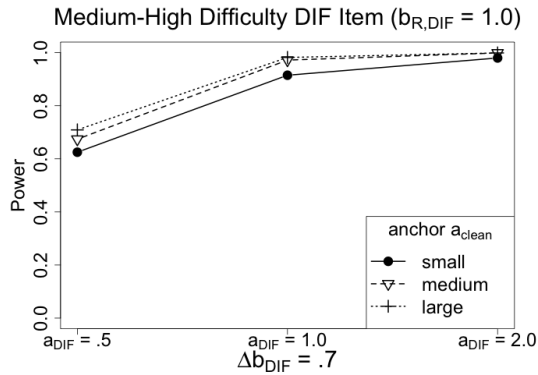


(c)

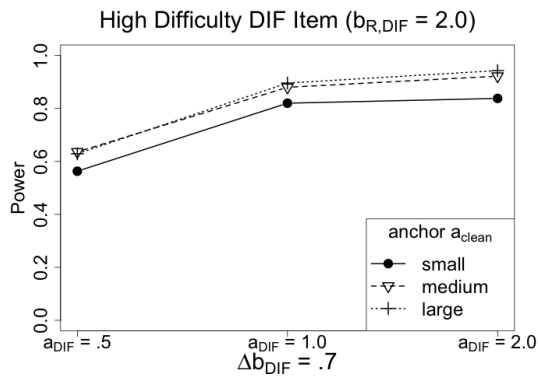
Figure 12:
Power rates across anchor configuration conditions for a negligible DIF effect size ($\Delta b_{DIF} = .3$) when (a) $b_{R,DIF} = 0$, (b) $b_{R,DIF} = 1.0$ and (c) $b_{R,DIF} = 2.0$.



(a)



(b)



(c)

Figure 13:
Power rates across anchor configuration conditions for a large DIF effect size ($\Delta b_{DIF} = .7$) when (a) $b_{R,DIF} = 0$, (b) $b_{R,DIF} = 1.0$ and (c) $b_{R,DIF} = 2.0$.