Detecting DIF in Forced-Choice Assessments:

A Simulation Study Examining the Effect of Model Misspecification

Contribution statement: Authors 1 and 4 developed the idea. Authors 1, 2, and 4 designed the simulation study. Author 1 wrote the code and analyzed the data. All authors contributed to the writing of the manuscript.

Open Science statement: All materials are available at https: <https://osf.io/fhd9w/?view_only=cccd50cce05a4dcea4df3a31fe963f2d>

**Abstract**

On a forced-choice (FC) questionnaire, the respondent must rank two or more items instead of indicating how much they agree with each of them. Research demonstrates that this format can reduce response bias. However, the data are ipsative, resulting in item scores that are not comparable across individuals. Advances in Item Response Theory have made scoring FC assessments possible, as well as evaluating their psychometric properties. These methodological developments have spurred increased use of FC assessments in applied educational, industrial, and psychological settings. Yet, a reliable method for testing differential item functioning (DIF), necessary for evaluating test bias, has not been established. In 2021, Lee and colleagues examined a latent-variable modelling approach for detecting DIF in forced-choice data and reported promising results. However, their research was focused on conditions where DIF items were known, which is not likely in practice. To build upon their work, we carried out a simulation study to evaluate the impact of model misspecification, using the Thurstonian-IRT model, on DIF detection, i.e., treating DIF items as non-DIF anchors. We manipulated the following factors: Sample size, whether the groups being tested for DIF had equal or unequal sample size, the number of traits, DIF effect size, the percentage of items with DIF, the analysis approach, the anchor set size, and the percent of DIF blocks in the anchor. Across 336 simulated conditions, we found [Results and discussion summarized here].

**Detecting DIF in Forced-Choice Assessments:**

**A Simulation Study Examining the Effects of Model Misspecifications**

Graphical user interface, text

Description automatically generated Forced-choice (FC) is a response format that requires respondents to arrange a set of items - known as a block - in the order that best represents them (see Fig. 1; Cao & Drasgow, 2019; Dueber et al., 2019). This technique was initially conceived to counteract construct-irrelevant variance by limiting respondents' capacity to endorse all desirable items (Bartram, 2007; Lee et al., 2018; Vasilopoulos et al., 2006). It is beneficial in high-stakes or sensitive assessment situations where desirable responding is more likely (Jackson et al. 2000). This includes educational and industrial settings where a test may help inform an admission or hiring decision, as well as the measurement of sensitive topics. For example, the Character Skills Snapshot is used in school admissions (Enrollment Management Association, 2023), the Mosaic (ACT, 2022) is used to inform program planning, and the Occupational Personality Questionnaire which is used in hiring decisions (Brown & Bartram, 2011). While a well-designed FC assessment can effectively reduce response bias, it produces ipsative data. Ipsative data occurs when the responses are directly dependent on each other: e.g., if you rank “I do not enjoy working in a group” first, then you necessarily must rank the other options in Figure 1. In contrast, a Likert style item allows for the selection of any response option, regardless of the response to the last item. FC assessments produce the same total score for each participant, making interindividual comparisons difficult. This type of data cannot be analyzed with standard methods. Advances in Item Response Theory (IRT) have enabled the evaluation and scoring of FC data, but methodologies for testing differential item functioning (DIF) still need development. Lee and colleagues (2021) introduced a latent variable modeling approach for DIF testing with FC data and evaluated the method using a simulation study. Their study showed promising results when anchor items were correctly specified. We expanded on this work by evaluating the method under conditions of realistic model misspecification when DIF items are included in the anchor set and different testing situations.

Figure 1. Forced-Choice Block Examples

This study aimed to further develop DIF testing methods for use in real-world FC testing scenarios. We begin with a review of IRT models and how they have been expanded to account for the complexities of FC data. We then explain the Thurstonian-IRT model and review the literature on methods for testing DIF with FC data to contextualize the current simulation study. Next, we state the research questions we sought to answer with this study and explain the various simulation conditions. Finally, we discuss the results and implications of this work.

**Modeling Forced-Choice Data**

Modern forced-choice models are based on IRT. IRT is a class of models used to understand the relationship between item responses and latent traits (Embertson & Reise, 2000). Most of these models assume local independence of items, meaning that after accounting for the factor, the items should be uncorrelated. This is a feature of traditional assessments where item A does not directly influence the score on item B after controlling for the latent trait. In FC assessments, however, the responses to each item within the same block are interdependent (the item-level data are ipsative; Baron, 1996), which also leads to the ipsative trait scores when traditional sum-score scoring approaches are used. With ipsative data, it is not possible to compare the scores of different individuals as each trait score is only comparable within an individual's responses. Brown and Maydeu-Olivares (2011) developed the Thurstonian-IRT model to address these challenges.

**The Thurstionian-IRT Model**

The Thurstonian Item Response Theory (TIRT) model is a normal ogive model with some special features. It is conceptually grounded in Thurstone's (1927) law of comparative judgment. According to this law, item *j* is preferred over item *k* in a pairwise comparison () if the latent utility of item *j* () is higher than that of *k* () (Brown & Maydeu-Olivares, 2011). This can be represented as:

Here, denotes the difference in latent utilities. When the differences in utilities yields a value below 0, item *k* is preferred over item *j*, and will equal 0.

***Model Specification***

The TIRT model can be specified as a second-order (Figure 2) or first-order model (Figure 3). The first-order model is primarily used for generating factor scores for participants. It involves estimating the thresholds and loadings from the pairwise comparison directly. In the second-order model, the observed responses are functions of the item utilities as described in Equation 1, and the utilities in turn are a linear function of the latent traits as described in Equation 2, where the utility equals the sum of the item mean (µ), the product of the loading (λ) and the measured traits (; expressed as a factor score), and an error term (ε) that is independent of the other items.

Figure 2 is a diagram for a simple second-order FC model consisting of three blocks with three items in each, measuring three traits. In this diagram, it can be seen that the pairwise item responses are a function of utilities which are then a function of the latent trait. This makes it a second-order model.

In the first-order model, the specification is reformulated in terms of which item will be preferred in a pairwise comparison between the items in equations 2 and 3. The differences between the traits are:

Furthermore, the difference between item means is often reduced to the threshold parameter:

The first-order model specification can be seen in Figure 3.

The main difference between the second-order equations in 2/3 and the first-order equation in 5 is that in the first-order model, differences in parameters values are examined for DIF across groups () verse just a utility mean () in the second-order model. We focus on the second-order model here because it estimates a single parameter to test for DIF across groups, instead of a difference score.

***Model Identification***

For model identification purposes, the loadings of each outcome of pairwise comparison () on the utilities and follow a patterned matrix of fixed values (either 1 or -1) as Figure 2 shows. Each is estimated without error as Equation 1 shows. Additionally, each trait's variance is set to 1 and their means are set to 0. Also, to set the metric of the utilities, one utility uniqueness per block is fixed to 1. Finally, to identify the scale origin of the latent utilities, one item utility mean per block is set to 0 (for example, the first item in each block) and all thresholds of the pairwise comparisons are explicitly fixed to 0 (Maydeu-Olivares & Brown, 2010). When this model is extended to multiple groups in the context of DIF testing, all constraints mentioned extend to both groups. Additionally, all utility error variances are set to be one in both groups along with a subset of the utility means (*t*). The amount of utility means set to be equal will vary depending on the size of the anchor set.

***Model Estimation***

The TIRT model is estimated in three stages. In stage one, the thresholds and tetrachoric correlations of pairwise comparisons are calculated. In stage two, the model parameters are estimated using a limited information estimator, such as unweighted least squares (ULS; Brown & Maydeu-Olivares, 2011). The utility means are then estimated such that the mean squared error of the residuals is minimized. The loadings are estimated from the tetrachoric correlations, again minimizing the residuals. Optionally, in stage three, the latent trait scores for respondents are estimated using maximum a posteriori estimation (MAP), which can be implemented in Mplus.

Figure 2

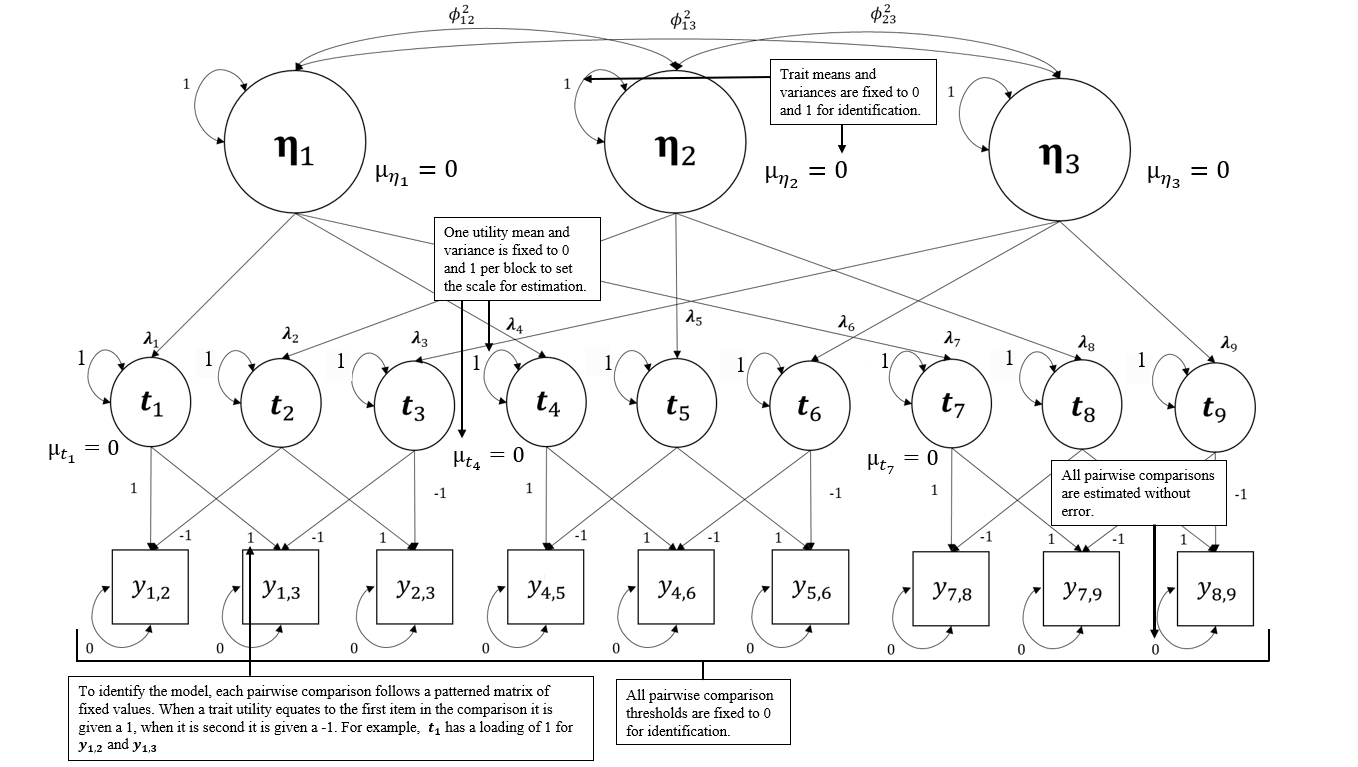
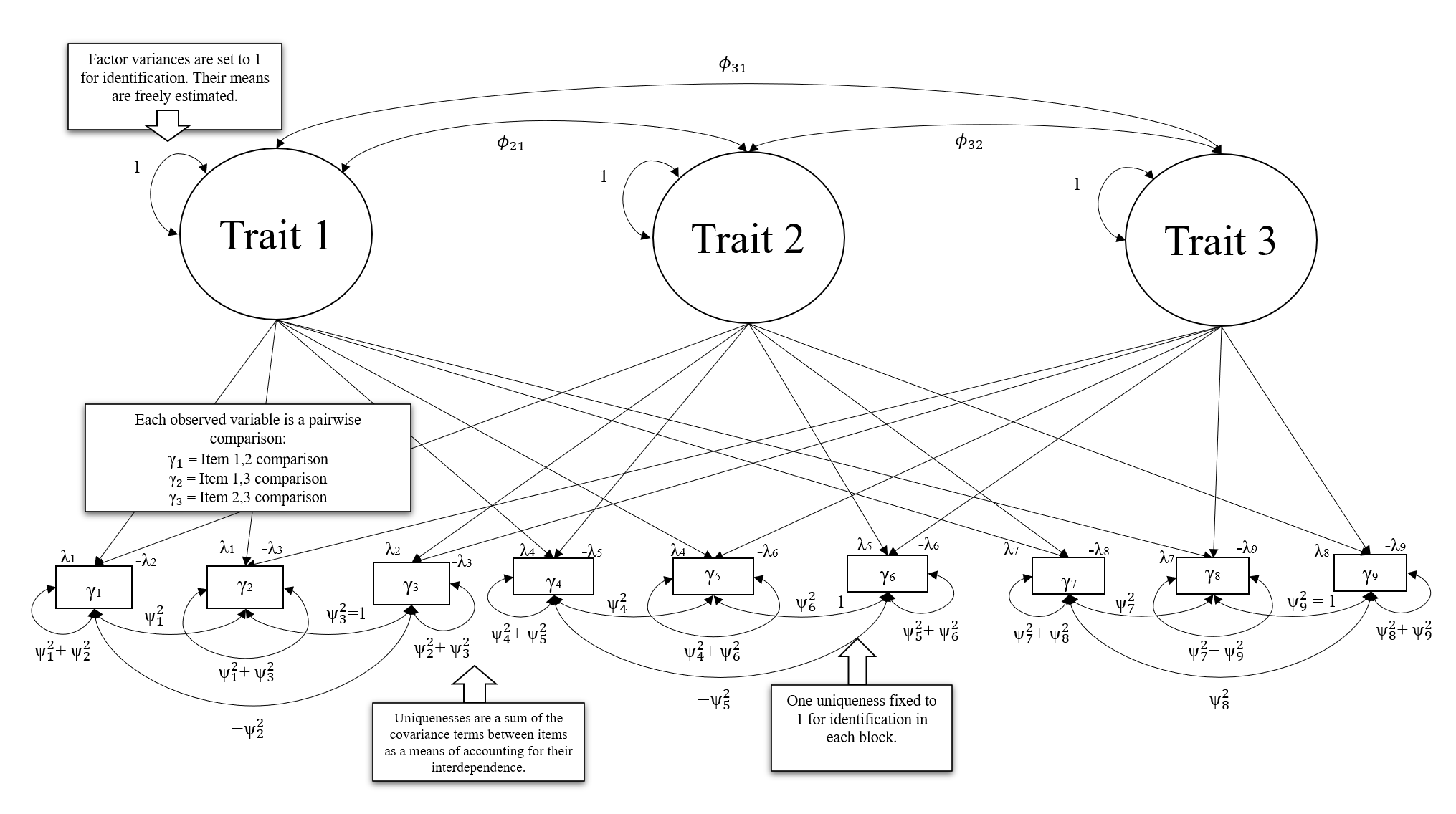
*Second-Order Thurstonian-IRT Model Diagram*

Figure 3

*First-Order Thurstonian-IRT Model Diagram*

**

**Differential Item Functioning and Forced-Choice Data**

The purpose of an assessment is to provide a valid score of examinees' abilities (AERA, 2014). Assessment items should not confer any advantage to specific groups based on factors unrelated to the measured construct, such as gender, race, or cultural background. When a group consistently scores higher than another group, this could be due to true higher ability (i.e., impact) or item bias (Dorans & Holland, 1992). DIF analyses can be used to identify differences between groups, other factors, or continuous data. In this study we examined DIF between groups. In this case DIF screens for potential bias by assessing if items perform differently across groups when their overall ability levels are equal (Angoff, 1993). DIF is classified into two categories: uniform and nonuniform.

Uniform DIF describes a consistent difference in the item's difficulty between two groups at all levels of the measured underlying trait (Swaminathan & Rogers, 1990). This suggests that the item is uniformly more or less difficult for one group compared to the other, irrespective of their standing on the trait. Factors such as biased item content that persistently impacts one group more than the other may cause uniform DIF (Camilli & Shepard, 1994). In the second-order model, uniform DIF is described by differences in the utility means (*t*) for two groups. Non-cognitive assessments don’t have correct or incorrect items, but items can still vary on difficulty in that they are harder to endorse, i.e., more of the latent trait is needed to endorse an item. Nonuniform DIF is characterized by differences in group performance on an item only occurring at some ability levels (Swaminathan & Rogers, 1990). This means that the item's slope for different groups varies depending on their trait level. This is each utility's loading (λ) on the latent trait in the TIRT model.

With FC data, DIF does not occur for a single item in isolation. This is because the items within a block are dependent on one another and responded to in a set, creating multidimensionality. However, most methods assume unidimensionality, such as Mantel-Haenszel, Delta-Plot, standardization, and logistic regression (Angoff, 1972; Dorans & Holland, 1992; Holland & Thayer, 1988; Swaminathan & Rogers, 1990) and thus are not appropriate for FC data. Instead, latent approaches, such as IRT models, that can handle multidimensional items are better suited for FC data.

**Item Response Theory Models for DIF**

The process of testing DIF in a latent-scoring approach typically involves specifying an IRT model where loading and threshold parameters are estimated for each group, except for a subset of anchor items constrained to be equal across groups. The remaining free parameters are tested to determine whether they differ significantly across groups. If they do, the item is flagged as displaying DIF. Various strategies for DIF testing have been proposed, including the free-baseline, constrained-baseline, and sequential free-baseline approaches which use model fit and/or Wald tests to determine the significance of differences (Stark et al., 2006).

The free-baseline approach requires constraining a subset of anchor indicators to be equal across groups, which establishes a comparison benchmark. The free-baseline approach is conducted over three steps. First, a set of items are constrained to be equal across groups, the anchor, while all others are freely estimated. Those freely estimated parameters are then tested for DIF using a multiple-constraint Wald test to determine if they are significantly different across groups. The multiple-constraint Wald-Test evaluates whether a set of coefficients in a model are equal, typically specified to test a difference of 0. The test statistic is calculated based on the estimated coefficients and their variance-covariance matrix, with larger values indicating more evidence of DIF. If the Wald test results in a significant difference for an item, it is flagged as a DIF item. The free-baseline approach is advantageous because it requires only a single model to be run (Stark et al., 2006). However, the optimal method for selecting the anchor set is unclear, but there is some evidence the anchor can be small if the anchor is of high quality (Lopez-Rivas et al., 2009).

This is juxtaposed against the constrained-baseline method, which follows a similar, yet opposite process: all items but one are constrained to be equal across groups, and the items are iteratively tested (Wang, 2004). The single freely estimated item is tested for DIF with a Wald test. Then, the next item is tested by constraining all remaining parameters again. This process is repeated until every item has been tested. The constrained-baseline approach performs well when the effect of DIF on the model is not severe (Stark et al., 2006) and allows for every item to be tested, in contrast with the free-baseline method where anchors cannot be tested (Chun, 2016). This approach’s disadvantages include the need for multiple model testing and the potential biasing effects of model misspecification (i.e., DIF items are held equal). Finally, there is the sequential free-baseline approach where DIF testing is done in two phases. In phase one, non-DIF items are identified by using the constrained-baseline approach. Then, in phase two, the non-DIF items from phase one are used as anchors in a free-baseline run of the model. Any items from this second phase that display DIF are then flagged (Chun et al., 2016).

To modify these methods for FC data, Lee and colleagues (2021) proposed that when identifying a multi-group first-order TIRT model, entire blocks of items, the thresholds and loadings, should be constrained to be equal across groups rather than a single item. With the first-order TIRT model it is not possible to parse out uniform vs nonuniform DIF as all parameters within a block need to be constrained. However, when using the second-order model, uniform DIF can be tested by examining constraints on the utility means (*t*) in the block. Nonuniform DIF can be analyzed by examining the loadings (λ).In the constrained-baseline approach, all blocks are constrained to be equal except for one. In the free-baseline approach, only the anchor blocks are constrained, while the remaining blocks are estimated freely in both groups. The free blocks are then tested for DIF by conducting a Wald test with degrees of freedom equal to the number of constraints estimated in each block. For example, when testing uniform DIF in a triad block, there are two *df*, one for each freely estimated utility mean in the block. Because the utility means within each block are interconnected and the items are multidimensional, the Wald test examines both freely estimated utility means simultaneously. In Fig. 2, this means would be constrained to 0 for model identification, then and would be tested for equality across groups simultaneously. In this way, the model tests *differential block functioning* across the groups rather than differential item functioning.

**Previous Research on DIF Testing in Forced-Choice**

DIF testing for FC assessments is a nascent area of research, for which only a few proposals have been made. Lin and Brown (2017) used a variation of a free-baseline method to detect DIF in two parallel FC test forms, where they estimated measurement parameters freely in each sample, and compared them after performing some scale equating. Wetzel and colleagues (2017) used a variation of the constrained-baseline method to test for DIF in an applied study of narcissism, where they constrained all TIRT measurement parameters equal across groups and examined the model modification indices to find items violating these constraints. Neither of these studies, however, had the formal objective to propose a method of DIF detection in FC questionnaires and thus did not systematically examine the merits of either method in various conditions. Lee and colleagues (2021) have proposed a formal method for detecting DIF, which is the free-baseline approach discussed above. In their study, they focused on block-specific factors, such as the type of DIF, the number of DIF items in each block, DIF effect size, presence of impact (differences in the means of the items outside of bias), and the number of DIF blocks on the test. They also examined sample size effects and differences between the constrained-baseline and free-baseline approaches. Overall, their results indicated that DIF blocks were consistently correctly identified (>95% of the time) across all block-specific conditions, except when the sample size was large (N=2000) and had impact (an actual difference in group ability) in the free-baseline approach. In contrast, detecting DIF blocks was far less accurate in the constrained-baseline approach across all conditions.

Lee and colleagues' (2021) work is consistent with other findings that have used the free-baseline and constrained-baseline approach for standard IRT models. For example, Stark and colleagues (2006) tested the approaches with dichotomous data in the 2PL and polytomous models using the graded response model. Across most simulated DIF conditions, the free-baseline approach was more accurate than the constrained-baseline approach. Several others have found a similar pattern of results in various IRT models (Chun et al., 2016; Kim et al., 2012; Woods & Grimm, 2011). Other research about the approaches has indicated that increasing the number of anchors subsequently results in more accurate DIF detection (Wang, 2004; Wang & Yeh, 2003). However, Wang (2004) found that using DIF items as anchors biased the remaining items toward one group, resulting in less accurate detection rates.

**The Present Study**

In the current study, we expanded Lee and colleagues' (2021) work to incorporate test and analysis features that are typical in practice. First, they examined a three and five-factor model, with 30 and 60 items (10 or 20 blocks) total, respectively. However, it is typical for personnel or education FC assessments to measure more traits (e.g., the 32-trait Occupational Personality Questionnaire, Brown & Bartram, 2009). Second, Lee and colleagues did not test different anchor set sizes or the impact of DIF blocks being included in the anchor set. These model misspecifications are relevant to real-world testing scenarios for which anchor items are unknown. Finally, their simulation used the first-order TIRT model, which does not allow for a direct examination of utility means. Only the mean differences between pairwise comparisons can be analyzed in the first-order TIRT model. It is useful to isolate the means of each utility as this indicates uniform DIF, making the second-order TIRT model more beneficial for DIF analyses. This also allows for uniform DIF and nonuniform DIF to be assessed in a FC model, although we examined only uniform DIF in this study.

First, we tested if the results from conditions investigated by Lee and colleagues replicate when using the second-order TIRT model. Then, we generated different conditions of model misspecification to determine the effectiveness of the free-baseline and constrained-baseline approaches. Our study sought to answer eight research questions. The first four questions (RQ1-4) focus on the replication of Lee and colleagues' work in the second-order model, while the remaining questions (RQ5-8) build upon it.

RQ1. Does the trend of increasing the number of traits resulting in better DIF detection from Lee and colleagues’ work (2021) replicate in the second-order model?

RQ2. Do the findings of Lee and colleagues (2021), which indicate a pattern where an increase in DIF effect size leads to enhanced power and consistent Type I error rates in both small and large effect size conditions, replicate when examined within the second-order model?

RQ3. Do the findings of Lee and colleagues (2021), where Type I error rates and power remained consistent in increasingly larger sample sizes (500 – 1000 per group), replicate in the second-order model.

RQ4. Does the result of the free-baseline method being more accurate than the constrained-baseline method for DIF detection replicate in the second-order model?

RQ5. What effect does having an assessment with a higher percentage of DIF blocks have on the accuracy of DIF detection in the second-order model?

RQ6. How does anchor set size influence DIF detection in the free-baseline approach when using the second-order model?

RQ7. To what extent does including DIF items in the anchor set reduce the accuracy of detection in the free-baseline approach when using the second-order model?

RQ8. What effect does unequal sample size have on detecting DIF in the presence of model misspecification?

RQs 1-4 provide needed replication research of methodological research (Loman et al., 2022), as well as an expansion to consider larger numbers of traits that are more consistent with assessment practice. RQ5 considers the effect of the different amounts of DIF for overall DIF detection, consistent with real world testing where the amount of DIF varies across tests. RQs 6-7 test the free-baseline approach under conditions of varying anchor size (RQ6) and model misspecification (RQ7). This is more representative of real-world contexts where the anchor items are not known prior to conducting the analysis. RQs 6 and 7 do not apply to the constrained-baseline model because in those models all but one block is constrained meaning the anchor set includes all blocks, DIF and non-DIF. RQ8 examined what effect unequal sample size has on DIF detection. For an overview of our study design, the hypotheses related to each question, and the meaning of various outcomes, see Appendix 1.

**Methods**

The code to reproduce the simulation can be found in the supplementary materials on the OSF page (<https://osf.io/fhd9w/?view_only=cccd50cce05a4dcea4df3a31fe963f2d>) of this manuscript. In designing the simulation study, we tested 11 replications of the five-trait, small effect size, with 40% of the items having DIF condition. The sample size was set to 1000 with equal groups. The free-baseline models with a 20% anchor set and no DIF items in the anchor set were used for this testing. These replications were generated only to ensure that simulation function worked and is reproducible. Reviewers may test the reproducibility of the simulation by changing the value of *k.range* to 11 at the start of the *SimCod\_TraitSize\_1\_ES\_1\_PDIF\_1\_FB\_SS\_1.R* file and using only the 5f\_20A\_0DIF[Block13DIF/Block18NON DIF].inp files. This is further explained in the *ReadME.txt* file on this manuscript’s OSF page.

**Conditions**

There are several conditions that remained constant across all replications. We describe these first before detailing the manipulated conditions. Each research question, except RQ4, has an accompanying condition that was manipulated. These can be broken down into data generation and analysis conditions.

***Constant Conditions***

**Analysis and Sample Factors.** The number of replications for each condition is set to 500. All conditions were tested using the TIRT model with Wald tests.

**Assessment Factors.** The block size was set to three items (triads).The number of blocks and items remained constant within each number of traits condition. There were 20 triad blocks (60 items total) in the five-trait condition. In the ten-trait condition, there were 40 triad blocks (120 items total). To increase the ecological validity of the simulation some of the trait correlations were negative and others positive. To further support ecological validity these correlations were based on a meta-analysis of the correlations of the Big Five personality traits (neuroticism, extraversion, openness, agreeableness, and conscientiousness; Linden et al., 2010). In the case of the Big Five , neuroticism is negatively correlated with the other 4 (-.36, -.17, -.36, -.43). This decision was also based on Frick and colleagues (2021) who found that parameter recovery was better when factors correlations were positive and negative. We used the matrix of correlations reported by Linden and colleagues (2010) in the five-trait condition and followed its pattern in the ten-trait condition. This meant that two traits were negatively correlated with all other traits and positively correlated with each other in the ten-trait condition. This was accomplished by randomly drawing absolute values from an inverse Wishart distribution with 100 degrees of freedom and covariances set to .3 and then making Traits 1 and 6 negatively correlated with the rest of the traits.

Including negatively keyed items, expressed as a negative factor loading, is important for accurately estimating the TIRT model (Brown & Maydeu-Olivares, 2011). In practice, the number of negatively keyed items would be kept low to avoid blocks having a clear 'best' (or desirable) answer. It has been shown that only 25% of blocks need to contain a negatively keyed item, and this is the amount we used (Lee et al., 2022). We simulated five blocks with negatively keyed items in the five-trait condition and ten in the ten-trait condition.

**Block Factors.** We varied DIF only at the utility mean (uniform DIF) level to focus the simulation on misspecification and sample size, while keeping it feasible. We only examined uniform DIF in this study. Previous work suggests that when uniform DIF was present, nonuniform DIF also tended to be detected (Lee et al., 2021). We also chose not to include nonuniform DIF as it would substantially increase the complexity of the study. We also only simulated one item per block with DIF. Lee and Colleagues (2021) considered multiple items per block with DIF with varied results. This was an area of complexity we chose not to introduce into the study. These conditions are summarized in Table 1.

Table 1

*Constant Simulation Conditions*

|  |  |  |  |
| --- | --- | --- | --- |
| Factor Type | | Factor | Levels |
| Analysis and Sample Factors |  | Replications  Model for Estimation  Parameter Test | 500  Thurstonian-IRT  Wald Test (2 *df*) |
| Assessment Factors |  | Number of blocks  Block size  Negatively Keyed  Blocks  Trait Correlations | 5 Trait = 20  10 Trait = 40  3  5 Trait = 5  10 Trait = 10  Positive and Negative  Correlations |
| Block Factors |  | Number of items with DIF in each block  Type of DIF | 1  Uniform |

***Manipulated Condition: Data Generation***

**Sample Size and Equality (RQ3/RQ7).** We simulated data such that there were either 1000 or 2000 total responses, just as Lee and colleagues (2021). In the authors’ collective experience working on FC assessments operationally, having sample sizes of 1000 or more is typical. We also expanded on Lee and colleagues by including an equal and unequal sample size condition. In real-world settings, equivalent groups are not often observed. When the sample sizes were equal the responses were evenly split (500/500 or 1000/1000 in each group. When sample sizes were unequal there were 25% more responses in one group (250/750 for the 1000 condition, 500/1500 for the 2000 condition).

**Number of Traits (RQ1).** We generated scores for respondents on five and ten traits. A five-trait analysis appeared in Lee and colleagues' (2021) original work and is also a common condition in other FC simulation studies (Frick et al., 2021; Schulte et al., 2021). Five-traits are assessed in the Big-Five FC assessment (Brown & Maydeu-Olivares, 2011). We also include a ten-trait condition to represent assessments used operationally such as the Character Skills Snapshot (seven traits; EMA, 2023) or OPAQ-32 (32-traits; Brown & Bartram, 2011). The number of items per trait was 12, resulting in 60 items (or 20 triad blocks) in the five-trait condition and 120 (40 blocks) in the 10-trait condition.

**DIF Effect Size (RQ2).** We simulated uniform DIF and manipulated the effect size of DIF by changing the amount added to the mean of items in one group that we selected to display DIF. Items were manipulated such that one group was consistently different on all DIF items (e.g., group 2 will always have effect size added to all items that display DIF). In the unequal sample size condition, the DIF effect size was always added to the smaller sample size group. For the magnitude of DIF, we rely on prior simulation research to determine the values because there has not been a practical examination of DIF for forced choice assessments. We used the same effect sizes from Lee and colleagues (2021) with a small effect size condition, .3, and a large one, .6. These effect sizes are in standardized units. These conditions are generally accepted as small and large in other simulations (Kim et al., 2016; Stark et al., 2006).

**Percent of DIF Blocks (RQ4).** The effect of the number of DIF blocks on DIF detection was examined by manipulating the percentage of items with DIF. When a block contained an item with DIF, we considered it a DIF block. We tested if there was a difference in the accuracy of DIF detection when 40%, 50%, or 60% of blocks display DIF. In the five-trait condition, this equated to 8, 10, or 12 total items with DIF. In the 10-trait condition, there were 16, 20, or 24 items with DIF.

***Manipulated Conditions: Analysis***

In addition to considering how different data features influence DIF detection, we tested different analysis features. In practice, a researcher will not know which blocks contain DIF prior to determining an anchor, and these conditions represent the different, yet reasonable, decisions researchers can make when testing for DIF.. The following conditions represent how the data were analyzed.

**Anchor Set Size (RQ5).** We examined the effect of anchor set size by manipulating the percentage of blocks specified as the anchors. Lee and colleagues (2021) did not study this beyond determining the minimum amount (20%) for model convergence in their study. We tested anchor block sets of 20% and 30%. For example, in the five-trait 20% condition this meant that four of the blocks are constrained equal across groups.

**Model-Misspecification (RQ6).** In this study a model misspecification is when a DIF block is used in the anchor set. We manipulated the amount of misspecification for the free-baseline conditions by varying the percent of DIF blocks included in the anchor set. 0%, 50%, or 100% of the total anchor blocks will have DIF. For example, in the 20% anchor set condition for five-traits there were four blocks in the anchor. In the 50% DIF in anchor set condition, two of these blocks contained DIF. There is little existing research from which we can base our decisions here, thus we tested a situation where the model was properly specified (0%), had some misspecification (50%), or was completely misspecified (100%). In all conditions, the percentage of blocks with DIF present remained constant regardless of the percent of DIF blocks in the anchor. For example, in the 40% blocks with DIF present condition for five traits, there were always eight DIF blocks on the test even as the number of DIF blocks in the anchor increased.

As the constrained-baseline approach constrains all but one block in each analysis, this condition and all other analysis factors are not applicable. This resulted in 288 conditions (2x2x2x2x3x2x3) for the free-baseline approach and 48 (2x2x2x2x3) conditions for the constrained-baseline approach. A summary of these conditions is in Table 2.

Table 2

*Manipulated Simulation Conditions – Data Generation*

|  |  |  |  |
| --- | --- | --- | --- |
| Factor Type | | Factor | Levels |
| **Data Generation Conditions** | | | |
| Sample Factors |  | Sample Size (Total)  Sample Size Equality | 1000, 2000  No/Yes  If No:  250/750  500/1500 |
| Assessment Factors |  | Number of Traits | 5, 10 |
| Block Factors |  | DIF Effect Size  Percent of Blocks With DIF | Small (.3), Large (.6)  40%, 50%, 60% |
| **Analysis Conditions** | | | |
| Modelling |  | Anchor set size (% of blocks)  DIF Included in Anchor  Latent Scoring Approach | 20%, 30%  0%, 50%, 100% (of total anchor set)  Free-Baseline or  Constrained- Baseline |

**Data Generation**

Data was generated in two phases in R using a modified function from Frick and colleagues (2021). In phase one the parameters were generated. This included:

1. Vectors of trait scores equal to the number of traits in the condition were simulated from a multivariate normal distribution, *MVN* (0,1), for each respondent in the reference and focal group. Trait correlations followed the approach described in the constant factors.
2. Vectors of loadings (λ) equal to the number of items in the condition were sampled from a uniform distribution, *U* (0.65, 0.95).
3. Vectors of item means (µ) equal to the number of items in the condition were sampled from a uniform distribution, *U* (-1,1) respectively. These are common values to use for continuous item utilities (Brown & Maydeu-Olivares, 2011). When DIF is present, the mean of one group was manipulated by adding the effect size to only one of the item means in the block.
4. Item errors (ε) were set to 1 in line with constraints placed on the model.
5. Measurement errors for each person were generated from a normal distribution, *N*(1, sqrt(ε)) for each item response.

In phase two, the generated variables and parameters were used to compute each pairwise comparison between two items in a block using the expression in equation 4. For example, in block one, the value for the pairwise comparison between items 1 and 2 was 1 if the linear combination of values sampled in step one, resulted in a value greater than or equal to 0 (indicating item 1 is preferred to item 2) and was 0 otherwise.

The simulation followed a two-step process. First, the datasets were generated in line with the data generation conditions in Table 2. Then they were subjected to each analysis condition. For example, 500 datasets for the five-trait, N =1000 equal groups condition with a DIF effect size of .3 added to 40% of the blocks were generated. Then, we analyzed them using the free-baseline model for the 20% anchor set with 0% DIF in the anchor condition. Followed by the 20% anchor set with 50% DIF in anchor, and so on until they were subjected to each analysis condition.

**Analysis Plan**

Analyses were conducted in R and Mplus. We used R to simulate the data, Mplus to analyze it (note that the TIRT model can be estimated in R using the thurstonianIRT package), and R to process the simulation results. The code to run the simulation, Mplus model files, and an analysis file to test our hypotheses are on the OSF (<https://osf.io/fhd9w/?view_only=cccd50cce05a4dcea4df3a31fe963f2d>)..

*Simulation Check*

We checked all replications for convergence via the calculation of the Wald Test, which will not be computed if the standard errors of parameters are not estimated. We denote these cases with ‘9999’ within the datasets in the Results folder. When non-convergence occurred, we checked those replications for errors and anomalies. There were a total of XX. instances of non-convergence.

***Model Estimation***

Models were analyzed in Mplus using the *MplusAutomation* package (Hallquist, 2022) in R. Every model was set up per the constraints and parameters described in the Thurstonian-IRT section. The number of item constraints across groups varied based on the anchor item percentage condition. In the 20% anchor set condition, four or eight blocks were constrained equal. This means that all item parameters in the block were constrained such that the utility means and loadings were set to be equal across groups. In the 30% condition, six or twelve blocks were constrained equal. Depending on the 'DIF included in anchor condition,' some number of blocks displaying DIF were included in the anchor set.

***Hypothesis Testing***

To answer our research questions and test our hypotheses, we examined if blocks containing a DIF item were accurately identified (indicated by a significant Wald test for each block). This was done by testing a subset of the freely estimated blocks in each condition. We tested four blocks, two with DIF and two without. This was done to reduce computation times and ensure parity in the number of statistical tests in each condition. Each Wald test was conducted with two degrees of freedom on the unconstrained utility means () in the block. Using the Wald test results, we calculated Type I error (α) rates as the proportion of non-DIF blocks incorrectly flagged as displaying DIF across replications and power (β) as 1 - the proportion of DIF blocks not flagged as DIF across replications.

**Results**

[Example results table are included below. We will also describe the results in the text.]

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Table 3.*  *Power and Type I Error in the Free-Baseline Conditions* | | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  | 5 Factor | | | | |  | 10 Factor | | | | |  |
|  |  | Percent DIF Items | Anchor Set Size | | DIF Included in Anchor | | Small DIF | |  | Large DIF | |  | Small DIF | |  | Large DIF | |  |
| Sample Size | Equal? | α | β |  | α | β |  | α | β |  | α | β |  |
| 1000 | Yes | 20 | 20 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 30 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 40 | 20 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 30 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 60 | 20 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 30 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| No | 20 | 20 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 30 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 40 | 20 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 30 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 60 | 20 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 30 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 2000 | Yes | 20 | 20 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 30 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 40 | 20 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 30 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 60 | 20 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 30 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| No | 20 | 20 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 30 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 40 | 20 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 30 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 60 | 20 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 30 | | 0 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 50 | |  |  |  |  |  |  |  |  |  |  |  |  |
| 100 | |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Table 4.*  *Power and Type I Error in the Constrained-Baseline Conditions* | | | | | | | |  |  |  |  |  |  |  |
|  |  |  |  | 5 Factor | | | | |  | 10 Factor | | | | |
|  |  | Percent DIF Items | | Small DIF | |  | Large DIF | |  | Small DIF | |  | Large DIF | |
| Sample Size | Equal? | α | β |  | α | β |  | α | β |  | α | β |
| 1000 | Yes | 20 | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 40 | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 60 | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| No | 20 | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 40 | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 60 | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 2000 | Yes | 20 | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 40 | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 60 | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| No | 20 | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 40 | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| 60 | |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

**Discussion**

* Results have implications for the practical usage of the free-baseline and constrained-baseline latent scoring approaches to detecting DIF in FC assessments.
  + Suppose it is found that the free-baseline method has low Type 1 error rates and high power even in cases of extreme misspecification. In that case, there is support for its usage in applied settings where these conditions may also be present.
  + Suppose the free-baseline method is better than the constrained-baseline. In that case, there will be support for selecting it as the approach of choice. It is less computationally intensive and can be computed with fewer model misspecifications (i.e., constraining 6 out of 20 blocks rather than 19 out of 20 in the five-trait condition).
* This research is also relevant more broadly to the FC response format as it lacks a well-researched method for detecting DIF. This research works toward providing that method so researchers can confidently use the latent variable modeling approach.
* Limitations:
  + While our conclusions may extend to other real-world scenarios, they are limited to the specific simulation conditions we tested.
  + The current study only examined DIF between groups. These results may not extend to continuous variables or more than two groups.
  + We did not examine nonuniform DIF. It still needs to be determined if nonuniform DIF can be accurately detected in the second-order TIRT model using these approaches. To confirm this, it may be worth examining nonuniform DIF along with other new conditions in a follow up study.
  + Our selection of simulated conditions is not based on real world analyses, because there are not any available. We base our decisions on other similar simulation studies. As more work on DIF for FC enters the literature, it will be important to expand simulation research to be representative of data and effects seen in practice.

**References**

ACT. (2022). Mosaic by ACT Social Emotional Learning Assessment Technical Manual: Elementary School Assessment. https://www.act.org/content/dam/act/unsecured/documents/2022/R2138-Mosaic-SEL-ES-Technical-Manual-03-2022.pdf

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). Standards for Educational and Psychological Testing. American Educational Research Association.

Baron, H. (1996). Strengths and limitations of ipsative measurement. Journal of Occupational and Organizational Psychology, 69(1), 49–56. https://doi.org/10.1111/j.2044-8325.1996.tb00599.x

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. International Journal of Selection and Assessment, 15(3), 263–272. https://doi.org/10.1111/j.1468-2389.2007.00386.x

Brown, A., & Bartram, D. (2009). Doing less but getting more: Improving forced-choice measures with IRT. Society for Industrial & Organizational Psychology annual conference, New Orleans. <https://kar.kent.ac.uk/44788/>

Brown, A., & Bartram, D. (2011). The Occupational Personality Questionnaire Revolution: Applying Item Response Theory to. *Organisationalpsychology.nz*. https://www.organisationalpsychology.nz/\_content/The\_Occupational\_Personality\_Questionnaire\_Revolution.pdf

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. Educational and Psychological Measurement, 71(3), 460–502.Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items (Vol. 4). Sage.

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. The Journal of Applied Psychology, 104(11), 1347–1368. <https://doi.org/10.1037/apl0000414>

Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016). MIMIC methods for detecting DIF among multiple groups: Exploring a new sequential-free baseline procedure. *Applied psychological measurement*, *40*(7), 486-499.

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel‐Haenszel and standardization 1, 2. ETS Research Report Series, 1992(1), i-40.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. Journal of Educational Measurement, 23(4), 355–368. <https://doi.org/10.1111/j.1745-3984.1986.tb00255.x>

Enrollment Management Association. (2023). Summary of the Snapshot research and findings. https://assets-global.website-files.com/62b5ff6837362e413f2bfed4/64e26e4744c35d763506450b\_summary-of-the-snapshot-research-and-findings.pdf

Frick, S., Brown, A., & Wetzel, E. (2021). Investigating the normativity of trait estimates from multidimensional forced-choice data. Multivariate Behavioral Research, 1-29.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the ETS Research Report Series, 1986(2), i–24. https://doi.org/10.1002/j.2330-8516.1986.tb00186.x

Jurgensen, C. E. (1944). Report on the "Classification Inventory," a personality test for industrial use. The Journal of Applied Psychology, 28(6), 445–460. https://doi.org/10.1037/h0053595

Kim, E. S., Joo, S. H., Lee, P., Wang, Y., & Stark, S. (2016). Measurement invariance testing across between-level latent classes using multilevel factor mixture modeling. Structural Equation Modeling: A Multidisciplinary Journal, 23(6), 870-887.

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC: Likelihood ratio test with a critical value adjustment. Educational and Psychological Measurement, 72(3), 469-492.

Kirkpatrick, J. J. (1951). Cross-validation of a forced-choice personality inventory. The Journal of Applied Psychology, 35(6), 413–417. <https://doi.org/10.1037/h0061581>

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22-56.

Lee, P., Joo, S.-H., & Stark, S. (2021). Detecting DIF in multidimensional forced choice measures using the Thurstonian Item Response Theory Model. Organizational Research Methods, 24(4), 739–771.

Lee, P., Joo, S.H., Zhou, S., & Son, M. (2022). Investigating the impact of negatively keyed statements on multidimensional forced-choice personality measures: A comparison of partially ipsative and IRT scoring methods. Personality and Individual Differences, p. 191, 111555. https://doi.org/10.1016/j.paid.2022.111555

Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced-choice measures with different scoring approaches. Personality and Individual Differences, 123, 229–235. <https://doi.org/10.1016/j.paid.2017.11.031>

Lin, Y., & Brown, A. (2017). Influence of Context on Item Parameters in Forced-Choice Personality Assessments. Educational and Psychological Measurement, 77(3), 389-414. DOI: 10.1177/0013164416646162.

Lohmann, A., Astivia, O. L., Morris, T. P., & Groenwold, R. H. (2022). It's time! Ten reasons to start replicating simulation studies. Frontiers in Epidemiology, 2, 973470.

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.

Maydeu-Olivares & Brown (2010). Item Response Modeling of Paired Comparison and Ranking Data. Multivariate Behavioral Research, 45: 935–974, DOI: 10.1080/00273171.2010.531231

Mazor, K. M., Clauser, B. E., & Hambleton, R. K. (1994). Identification of nonuniform differential item functioning using a variation of the Mantel-Haenszel procedure. Educational and Psychological Measurement, 54(2), 284–291.

Richardson, M. W., & Kuder, G. F. (1933). Making a rating scale that measures. The Personnel Journal, pp. 12, 36–40. https://psycnet.apa.org/fulltext/1933-04735-001.pdf

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. The Journal of Applied Psychology, 91(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Thurstone, L. L. (2017). A law of comparative judgment. In Scaling (pp. 81–92). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315128948-7/law-comparative-judgment-louis-thurstone

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Human Performance, 19(3), 175–199. <https://doi.org/10.1207/s15327043hup1903_1>

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.

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. Applied Psychological Measurement, 27(6), 479-498.

Wetzel, E., Brown, A., Hill, P., Chung, J., Robins, R., & Roberts, B. (2017). The narcissism epidemic is dead; long live the narcissism epidemic. Psychological Science, 28(12), 1833-1847.

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.

|  |  |  |  |
| --- | --- | --- | --- |
| **Appendix 1.**  *Interpretations given different outcomes.* | | | |
| **Research Question** | **Hypotheses** | **Analysis Plan** | **Interpretations given different outcomes** |
| 1. Does the trend of increasing trait size resulting in better DIF detection from Lee and colleagues (2021) replicate in the second-order model? | 1. When an assessment has ten traits, non-DIF items will be less likely to be incorrectly flagged as displaying DIF (lower Type I error rates) compared to the five-trait condition. 2. When an assessment has ten traits, DIF items will be more likely to be correctly identified as displaying DIF (higher power) compared to the five-trait condition. | A comparative table will be used to show the proportion of Type I error rates and the power in each condition. The table will separate the results by method. See Table 3 for the free-baseline approach and Table 4 for the constrained-baseline approach. | If our hypotheses are correct, there would be evidence that assessments containing more traits have a greater ability to detect DIF items. This aligns with previous studies that have shown that assessments with more dimensions tend to have increased normativity (Schulte et al., 2021). Because of this increased normativity, it should be easier to detect DIF items, as they are less influenced by the ipsative-specific characteristics of FC assessments.  If we find that trait size does not have an effect, there is evidence that assessments can be designed more intentionally to measure only what is essential, rather than incorporating additional traits to enhance DIF detection.  If a smaller trait size results in better power and lower Type I error rates, this could be an important consideration for researchers when designing their assessments.  In the event that there is reduced power (indicating hypothesis B isn't upheld) and lower Type I error rates (indicating hypothesis A is supported) in the five-trait vs. ten-trait condition, this suggests that increasing the number of traits might decrease the likelihood of false positives but heighten the risk of undetected DIF. |
| 1. Do the findings of Lee and colleagues (2021), which indicate a pattern where an increase in DIF effect size leads to enhanced power and consistent Type I error rates in both small and large effect size conditions, replicate when examined within a second-order model? | 1. A higher proportion of DIF blocks will be correctly flagged as DIF with a greater DIF effect size (indicating greater power). 2. The proportion of non-DIF blocks incorrectly flagged as displaying DIF will remain consistent (within .01 units of each other) across different effect size conditions. | A comparative table will be used to show the proportion of Type I error rates and the power in each condition. The table will separate the results by method. See Table 3 for the free-baseline approach and Table 4 for the constrained-baseline approach. | If our results replicate those of Lee and colleagues (2021), and a higher proportion of DIF blocks are correctly flagged as DIF when the DIF effect size is larger, and the proportion of non-DIF blocks incorrectly flagged as displaying DIF remains consistent, then there would be further evidence for the latent scoring approach being usable under various conditions of DIF being present.  If we find results that diverge from prior findings, there would be evidence for potentially needing a different method depending on the anticipated magnitude of DIF.  If hypothesis A is supported, there's evidence that the methods more accurately detect non-DIF blocks when its effect is larger in the second-order model. If hypothesis A is not supported, it means that, irrespective of effect size, the methods will flag non-DIF blocks with the same accuracy.  If hypothesis B is upheld, it means that both methods can consistently flag DIF blocks as DIF regardless of the effect size in the second-order model. If not supported, this will suggest the findings from Lee and colleagues (2021) don't mirror in the second-order model and that the accuracy of DIF detection varies.  If hypothesis B is supported but hypothesis A isn't, this would hint that as the DIF effect size grows, the methods' capacity to correctly flag non-DIF blocks might decrease, but the risk of not flagging DIF blocks remains unchanged. |
| 1. Do the findings of Lee and colleagues (2021), where Type I error rates and power remained consistent in increasingly larger sample sizes (500 – 1000 – 2000 per group) replicate in the second-order model? | 1. Sample size differences will not affect power in the second-order model (they will be within .01 units of each other in both conditions). 2. Sample size differences will not affect the Type I error rates in the second-order model (they will be within .01 units of each other in both conditions). | A comparative table will be used to show the proportion of Type I error rates and the power in each condition. The table will separate the results by method. See Table 3 for the free-baseline approach and Table 4 for the constrained-baseline approach. | If we observe that the proportion of DIF blocks correctly flagged as DIF remains consistent across varying sample sizes in the second-order model, and the differences are within .01 units, this would reinforce Lee and colleagues' (2021) findings. It would indicate that, similar to their study, power for DIF detection in the second-order model remains acceptable at a smaller or larger sample size.  If Type I error rates remain similar across different sample sizes in the second-order model it would be in line with the results observed by Lee and colleagues (2021). This would suggest that the ability of the methods in the second-order model to identify non-DIF items correctly remains constant. |
| 1. Does the result of the free-baseline method being more accurate than the constrained-baseline method for DIF detection replicate in the second-order model? | 1. The power of the free-baseline method will be higher than that of the constrained-baseline method. 2. The Type I error rates for the free-baseline method will be lower than that for the constrained-baseline method. | A comparative table will be used to show the proportion of Type I error rates and the power in each condition. The table will separate the results by method. See Table 3 for the free-baseline approach and Table 4 for the constrained-baseline approach. | If our results show that the free-baseline method correctly flags a higher proportion of DIF blocks as DIF (higher power) and incorrectly flags a lower proportion of non-DIF blocks as displaying DIF (lower Type I error rates) than the constrained-baseline method, further evidence will support its use over the constrained-baseline method.  However, if we find the inverse to be true, our findings will conflict with that of Lee and colleagues (2021), making it uncertain which latent scoring approach is superior.  If the free-baseline method incorrectly flags a lower proportion of non-DIF blocks as displaying DIF (lower Type I error rates) but also correctly flags a lower proportion of DIF blocks as DIF (lower power) compared to the constrained-baseline approach, this may suggest that it is better at avoiding false DIF flags but worse at correctly identifying actual DIF blocks. |
| 1. What effect does having an assessment with a higher percentage of DIF blocks have on the accuracy of DIF detection in the second-order model? | 1. As the number of DIF items on the test increases, power will decrease but remain acceptable. 2. As the number of DIF items on the test increases, the proportion of non-DIF blocks incorrectly flagged as displaying DIF will increase. | A comparative table will be used to show the proportion of Type I error rates and the power in each condition. The table will separate the results by method. See Table 3 for the free-baseline approach and Table 4 for the constrained-baseline approach. | If the Type I error rates remain constant or decrease, this will mean that regardless of the quantity of blocks with DIF on the test, non-DIF blocks are being correctly identified.  If power decreases as the number of DIF blocks increases, this will mean that the number of blocks with DIF impacts the ability to correctly detect those DIF blocks.  If the inverse is true and power increases, the increased quantity of blocks with DIF on the test improves the identification of DIF blocks. This would be a truly crazy outcome. |
| 1. How does anchor set size influence DIF detection in the free-baseline approach when using the second-order model? | 1. Increasing the anchor set size will improve the proportion of DIF blocks correctly flagged as DIF in the 0% DIF in anchor conditions (indicating improved power). 2. Increasing anchor set size in the 50% and 100% DIF in anchor conditions will lead to a higher proportion of non-DIF blocks incorrectly flagged as displaying DIF and a reduced proportion of DIF blocks correctly flagged as DIF (lower power and higher Type I error rates.) | A comparative table will be used to show the proportion of Type I error rates and the power in each condition. The table will separate the results by method. See Table 3 for the free-baseline approach and Table 4 for the constrained-baseline approach. | If we find evidence for the anchor set size resulting in a higher proportion of DIF blocks correctly flagged as DIF in the 0% DIF in anchor set condition, there would be support for using larger anchor sets when possible. This is consistent with the literature (Kopf et al., 2015).  However, if we also find that the proportion of DIF blocks correctly flagged as DIF remains constant regardless of anchor set size, this implies that researchers might be able to choose a smaller pure anchor set for their needs.  If we find support for hypothesis B. There would be evidence for using a smaller anchor set when the quality of the anchor set is unknown to enhance the proportion of DIF blocks correctly flagged as DIF.  Conversely, if the alternative to hypothesis B is supported, it suggests that using a larger anchor set, even when the quality of the anchor set is uncertain, may still maintain or even enhance the ability to correctly detect DIF blocks, without significantly increasing the risk of non-DIF blocks being incorrectly flagged as DIF. |
| 1. To what extent does including DIF items in the anchor set reduce the accuracy of detection in the free-baseline approach when using the second-order model? | 1. As the percentage of misspecification increases, the proportion of non-DIF blocks incorrectly flagged as displaying DIF will also increase (meaning Type I error decreases). 2. As the percentage of misspecification increases, the proportion of DIF blocks correctly flagged as DIF will decrease (power will decrease). 3. Although power will decrease and Type I error rates will increase as model misspecification increases, they will not move past an unacceptable threshold. Type I error rates will still be less than .05 and power will be above .8. | A comparative table will be used to show the proportion of Type I error rates and the power in each condition. The table will separate the results by method. See Table 3 for the free-baseline approach and Table 4 for the constrained-baseline approach. | If hypothesis A is supported, this will indicate that increasing levels of misspecification does affect the ability of the free-baseline method to detect DIF blocks correctly. If it is not supported, this will suggest that the method can detect DIF consistently, irrespective of the level of misspecification. This could mean that either misspecification doesn't impact DIF detection, or even a small level of misspecification compromises the method's ability to detect DIF.  If hypothesis B is supported, this will indicate that as model misspecification increases, the method's ability to accurately detect non-DIF blocks as not displaying DIF diminishes. Conversely, if it is not supported, it could mean that the degree of misspecification affects the method's precision in detecting non-DIF blocks. It might also suggest that even minimal misspecification can influence the ability to accurately detect non-DIF blocks.  Furthermore, if hypothesis C is supported, it will imply that DIF can be detected even in cases of significant misspecifications. This provides evidence that, regardless of anchor quality, test developers can still accurately detect DIF. If it is not supported, then as misspecification increases, it becomes more challenging to accurately detect DIF, echoing the sentiments of hypothesis B regarding power. |
| 1. What effect does unequal sample size have on detecting DIF in the presence of model misspecification | 1. Power will decrease with unequal sample sizes in the presence of model misspecification compared to equal sample sizes. 2. Unequal sample sizes will lead to an increase in Type I error rates when the model is misspecified compared to equal sample sizes. | A comparative table will be used to show the proportion of Type I error rates and the power in each condition. The table will separate the results by method. See Table 3 for the free-baseline approach and Table 4 for the constrained-baseline approach. | If power decreases (hypothesis A is supported) and there is a decline in the proportion of DIF blocks correctly flagged as DIF when the model is misspecified and sample sizes are unequal, this suggests that as the model increases in misspecification, imbalanced sample sizes reduce the ability to effectively detect genuine DIF blocks.  If we observe an increase in power when the samples are unequal in a misspecified model, it suggests an unexpected outcome where the disparity in sample sizes might be compensating for or highlighting model errors, aiding DIF detection.  If there is support for hypothesis B and we find that the proportion of non-DIF blocks incorrectly flagged as displaying DIF increases with unequal sample sizes in a misspecified model, this would indicate that model misspecification combined with sample size imbalances exacerbates the risk of false positives in DIF detection.  If the proportions of correctly flagged DIF blocks and incorrectly flagged non-DIF blocks remain relatively constant regardless of sample size variation in a misspecified model, it would imply unequal sample size in each group can be used without a reduction in the ability to detect DIF compared to equal sample sizes. |
| Interactions | 1. Anchor set size X Model misspecification. 2. Anchor set size X Blocks with DIF 3. Model misspecification X Block with DIF 4. Anchor set size X Model misspecification. X Block with DIF | An ANOVA with the model misspecification, anchor set size, and blocks with DIF variables entered as independent variables and either Type I error or Power entered as the dependent variable will be run to examine the interaction effects. When the *p* values are < .05, this will indicate a significant interaction. Additional post-hoc tests will be run to determine which groups are different. | 1. This interaction investigates how the combined effect of anchor set size and model misspecification influences DIF detection. If Type I error rates or power vary at different levels (with a threshold of .1 units) it suggests that the effect of the anchor set size on DIF detection may vary depending on the level of model misspecification. If it is found that regardless the level of model misspecification at different anchor set size, power and Type I remain constant and at acceptable level, there will be evidence that the method is applicable when model misspecification is large and the anchor set is small. 2. This interaction examines whether the influence of anchor set size on DIF detection varies depending on the number of blocks with DIF. An interaction would imply that the relationship between anchor set size and DIF detection changes as the number of blocks with DIF changes. This may indicate that regardless of the size of the anchor, if too much of the test is contaminated with DIF blocks, it will not be possible to accurately identify them or vice versa. 3. The interaction explores how model misspecification and the number of blocks with DIF interact to affect DIF detection. If there is evidence of this interaction, it would suggest that the effect of model misspecification on DIF detection is contingent on the number of blocks with DIF on the test. 4. This three-way interaction explores how the combined effects of anchor set size, model misspecification, and blocks with DIF interact to affect DIF detection. A significant interaction would suggest a complex interplay among these three factors in influencing DIF detection. |