Dear Dr. Montoya

We would first like to thank you for the opportunity to revise our manuscript and resubmit it. We also would like to thank you and the two reviewers for their constructive feedback. After reviewing the feedback, we identified two major themes arising from the comments.

1. Interest in adding nonuniform DIF and additional sample size conditions.
2. Several points that were unclear or confusing that needed to be rewritten.

We provide an overview of the changes we've made to address these issues here but also have responded to yours and the reviewers' comments below. Our responses are in **bold text** and we have included passages from the manuscript that have changed when necessary. These are highlighted**.** When applicable, we have indicated the page number of changes in our responses. These revisions were extensive and resulted in much of the manuscript being updated. As such we have uploaded a clean version of the document rather than a tracked version.

In response to the desire for a nonuniform DF condition, we believe that this would expand the scope of our manuscript, introduce much greater complexity, and increase computation times past our desired point. We agree with Dr. Montoya, that justification is warranted for not including them. As such we have added the following text in response to a lack of a nonuniform DIF condition (p. 19):

> 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 simulate one item per block with DIF. Lee and Colleagues (2021) considered multiple items per block with DIF with varied results.

In response to the point raised that a higher sample size condition should be examined we have added a 2000-person condition. We also agree that examining unequal sample sizes would improve the applicability of the simulation to real-world conditions. As such we have introduced an unequal/equal condition for each sample size. These conditions have new questions associated with them and have been added to Appendix 1. These conditions are described on p.22:

> We simulated data such that there were either 1000 or 2000 total responses. Lee and colleagues (2021) used these sample sizes in their study. We also expanded on Lee and colleagues by including an equal and unequal sample size condition. 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).

To point two, there were numerous points that needed clarified. We have rectified these issues throughout the manuscript. There are far too many to list here, however, each is listed in the response to reviewers below.

We again would like to thank you and the reviewers for your time and feedback, which has greatly improved the manuscript.

Editor

Thank you for your submission to PCI-RR! The submitted manuscript shows promise but needs additional revision prior to further consideration. We received comments from two very engaged reviewers, and I think their comments clearly identify some steps forward that should be considered. I want to emphasize a few of their comments and provide a few of my own which I do below.

Thank you for drawing our attention to these comments and your extensive feedback. It is appreciated!

Commentary on Reviewer Comments

I agree with Reviewer 1 that incorporating non-uniform DIF into the simulation would strengthen the contribution of the study. I understand this could be a large undertaking within the study, so it may not be feasible to extend the study in this dimension. Instead perhaps it would be worthwhile to comment on whether the researchers hypothesize the results to be similar or dissimilar for uniform vs. non-uniform DIF (i.e., whether you think it's safe to generalize the results to non-uniform DIF, or whether further studies may be needed to explore this phenomenon in non-uniform DIF contexts [or mixed contexts]).

Thank you very much for this point. We do not believe it is in the scope of this work to examine nonuniform DIF for the reason you mentioned. However, the results of Lee and Colleagues (2021) study indicates it may be possible to generalize our results of finding uniform DIF to the ability to find nonuniform DIF. However, it may also be worthwhile to examine it in future research as well. We have justified our decision to not include it in the Block Factors section. It reads (p. 19):

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 simulate one item per block with DIF. Lee and Colleagues (2021) considered multiple items per block with DIF with varied results.

Similarly, I found the Reviewer 1's comment on sample size (equal vs. unequal) to be quite interesting. It is common to test DIF across demographic characteristics which are often unbalanced. This may have some effect on the results of the study and should potentially be incorporated or discussed. Note that Reviewer 2 suggested testing multiple sample sizes or to cite literature which suggests how sample size might impact the outcome.

Thank you for drawing our attention to this point. Along with reviewer 2's feedback about a larger sample size condition, we have added a new sample size condition and an unequal/equal condition (p.19):

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). We also expanded on Lee and colleagues by including an equal and unequal sample size condition. 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).

I agree with Reviewer 2 that the term ipsative could be more clearly defined.

Aligned with their recommendation we have added a definition to the very start of the manuscript (p.3):
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 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.

A couple things to consider because this is a registered report: Review carefully the introduction and methods as these sections will not be able to be changed after Stage 1 IPA (other than light editing). Many parts of the manuscript are written in future tense, but other parts are in past tense. I recommend revising everything to past tense because you'll need to make these edits for Stage 2 anyway.

Thank you very much for the feedback. We have changed all wordings to past tense.

For your simulation you should consider some positive checks that you might conduct prior to doing your analysis, that would help evaluate that the simulation has been conducted correctly. These could be any evaluation of factors that might suggest failures of the data generation or data analysis process. In simulations this might include things like, reporting rates of convergence, checking parameter bias for conditions where bias should be zero, etc.

Great point, thank you. We have added a check that will record non-convergence in the code. We are conceptualizing this as when the standard errors cannot be calculated (thus a Wald test cannot be computed). We added a section to our analysis plan addressing this (p.24):
We checked all replications for convergence via the calculation of the Wald Test, which will not be computed if the modes parameters are not estimated. We denote these cases with '9999' within the datasets in the Results folder in . When non-convergence occurred, we checked those replications for errors and anomalies. There were a total of XX. instances of non-convergence.

Appendix 1. Design Table.

Please add an analysis plan column to the study design table describing how the research team will determine whether each hypothesis is supported. This could involve some test statistic or effect size, a visualization with a specific pattern, a table with a specific pattern. However, it should be clearly unambiguous how to interpret the result of the test, and not based on the judgement of the research.

Our apologies for missing this. We have added an Analysis Plan column and include the edited appendix in its entirety to address all of the Appendix 1 comments. We intended Tables 3 and 4 to be used for a visual comparison between the conditions and have made that clear in the analysis plan.

For each of your comments below we state in brief how we addressed them, however, we have expanded the table to include more interpretations and the additional questions from adding the sample size conditions. We have also added a row about interaction effects in line with reviewer 2. This updated appendix starts on p. 39.

In general, the interpretations column does not sufficiently account for all possible outcomes of the research results. Especially when there are multiple outcome variables and results could be mixed across these outcomes.

RQ1 should be broken into two hypotheses, as there are two outcomes. The interpretation should acknowledge the potential for contradictory results across outcomes and how such a result would be interpreted. For example, what if TIE is lower and power is also lower?

We broke up the hypothesis in two and added an interpretation for the outcome you mentioned.

As a note we also broke up RQ3 and 4's hypotheses as well to align with this feedback.

RQ2 it's unclear how Type I Error rates will be compared. The threshold is set to .01, but is this of the estimates from each condition, or based on some kind of uncertainty estimate. In addition, it's not clear if the second part of the interpretation is specific only to Type I Error or also power. These could be letters (a, b, or ab) to indicate which claims would be made based on the outcomes of which tests.

We have included additional interpretations for each hypothesis.

RQ4 This section I found very confusing. It seems that in this case a Type I Error would be when DIF is not present, but it is detected. But the interpretation suggests that if Type I Error is constant or decreases this suggests that DIF can be correctly identified, which sounds more like power than Type I error. Similar issues arise with the other interpretations, so perhaps clearly defining what is meant by a type I error in this case would be helpful.

We further clarified this to make it more intuitive in the hypothesis testing section (p.24).

RQ6. Please clarify what is an acceptable type I error rate and power. I found the second section of the interpretation difficult to understand.

We added what is considered acceptable into the hypotheses.

Introduction

Equation 1 is not fully defined because the value of y is not defined if y* < 0

Thank you for the note. We have added the second part of this equation on p.6:

$$y_{jk} = 0 \; if \; y^*_{jk} = t_j - t_k < 0$$

In Equation 2/3 is there an assumption that e_j and e_k are independent, or can they be dependent?

This is clarified now on p. 6:

… and an independent error term (ε).

In general I found it difficult to follow the language of "first order" and "second order" models. This likely stems from my lack of background in this area, but I think it could be more clearly explained for interested yet less embedded audience members.

Thank you for the note. We have added greater context in the model-specification section (p. 5-6) and added a diagram to show the first-order model (p.9):

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 (η; expressed as a factor score), and an error term (ε) that is independent of the other items.

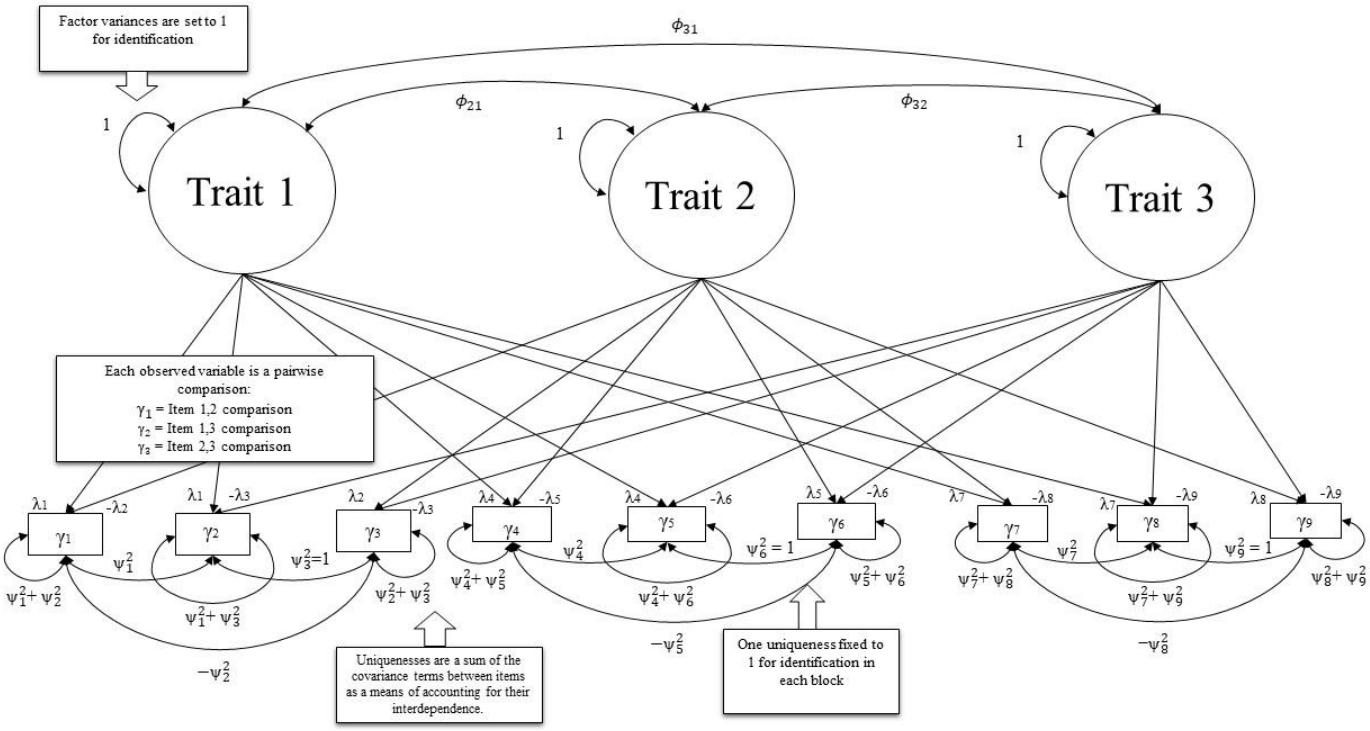
$$t_j = \mu_j + \lambda_j \eta_j + \varepsilon_j \quad (2)$$

$$t_k = \mu_k + \lambda_k \eta_k + \varepsilon_k \quad (3)$$

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.

Figure 3

First-Order Thurstonian-IRT Model Diagram



There is some discussion of the modeling distinction of DIF for an item vs. a pairwise comparison, but it was not really clear what the practical implications of this would be. Perhaps providing an example of when one vs. the other would be expected?

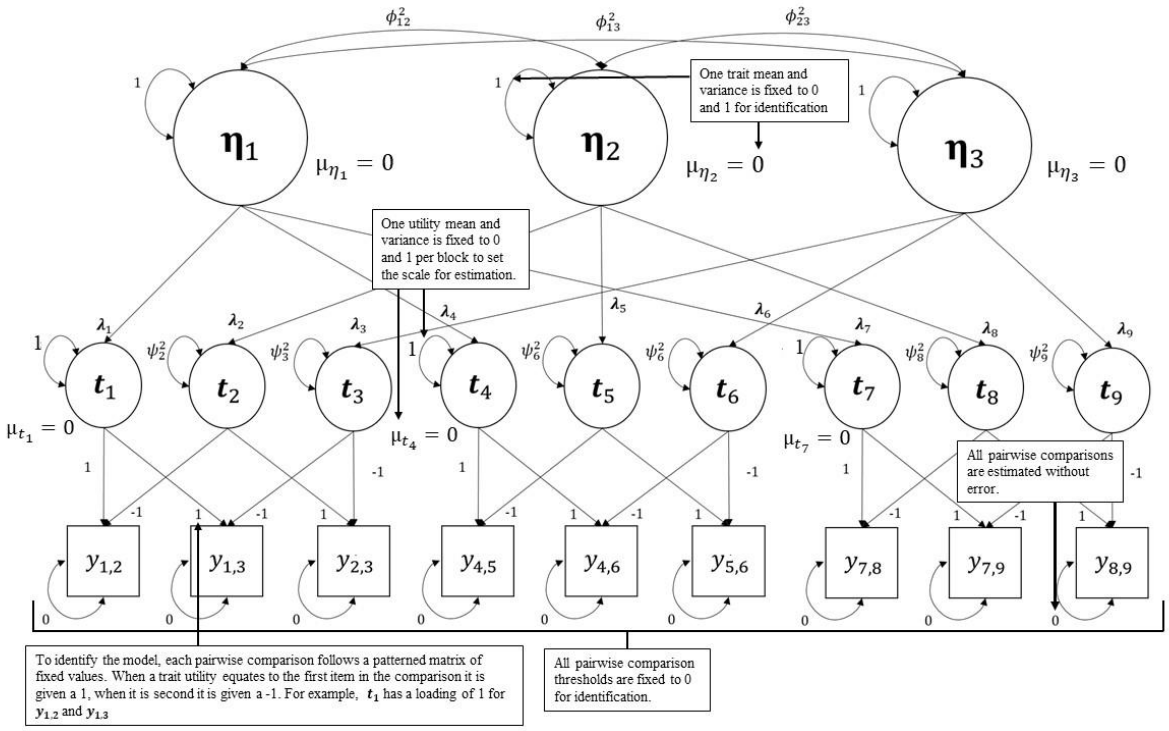
Thank you for the comment. We have clarified that we meant item utility in the section we believe you are referencing. We have also added additional context as to why it is generally desirable to be working with item utilities rather than pairwise comparisons for DIF analyses. This reads (p.6):

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 ($\mu_j - \mu_k$) verse just a utility mean (t_j) 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.

I found the Model Identification section to be quite difficult to read. Perhaps incorporating the notation which is introduced in the prior pages to clarify what is meant in this section would be

helpful. A reviewer also suggested incorporating some of this information in a figure, which is an idea I support.

Aligned with the reviewer's feedback, we have added additional information to Fig. 2 and included it below. We have also changed some of the words to just the notation from the equations introduced in the prior section.



While there are obviously a lot of abbreviations used throughout the paper, I think they could be avoided in section headers to make the contents more clear to someone who might be skimming the paper.

With the exception of the Previous Research on DIF Testing in Forced-Choice section we have removed all abbreviations from section headers. We changed FC to Forced-Choice in this one, but it would go to two lines if we changed DIF.

While DIF testing by groups is somewhat of a norm, DIF can be defined and tested by continuous variables. For example, using the MNLFA approach by Bauer. It is a limitation of the study that the exploration of DIF is only by groups, and this should be acknowledged. The definition of DIF and discussion of DIF seems to imply that it only applies to groups, which is not quite right. So differentiating the kind of DIF testing this paper is doing vs. what a broader definition of DIF is.

We added a bit of text to state there are other types of DIF analyses (p.10):

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

We added a new limitation on p.32 that we will include:

The current study only examined DIF between groups. These results may not extend to continuous variables or more than two groups.

I found the paragraph at the bottom of page 8 quite difficult to understand. I was confused because I thought the utility means had been fixed as part of the identification section, but maybe I'm not understanding which means are fixed. Similarly, the explanation of nonuniform DIF sounds backwards. Perhaps part of this is that I'm unfamiliar with using the term preferability. If preferability is the same as discrimination, would not differences on preferability correspond to uniform (rather that nonuniform DIF)? I do really struggle with the term preferability, because it seems reverse to difficulty. For example, options that many people endorse I would describe as highly preferable, but that would be low in difficulty. Again, this might be my out-of-field experience showing, but I didn't find the language intuitive, and a cursory search did not turn up any papers using this term.

Thank you for pointing out the lack of clarity here. We wanted to ground the terminology in non-cognitive assessment. As such, items are just more preferred or less preferred. However, we recognize this can be unintuitive. As such, we have changed the terminology to be consistent with how it is typically discussed. This section now reads (p. 10):

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.

In general, I found it difficult to determine whether the IRT Models for DIF section was describing the performance of the models based on prior research, or rather whether the claims were speculative. There are very few citations in this section, suggesting perhaps there is little research in this area. But the claims seem somewhat strong. It should be more clear if these claims are merely based on the researcher's hypotheses, or whether they are founded in prior research. Additionally, as a smaller note I found myself wondering if there are any advantages of a constrained-baseline approach. Only disadvantages were mentioned.

Thank you for the note. We have added more references throughout the section and also taken out some of the language that was too strong, and that we could not support

(specifically for the Sequential Free-Baseline method). We also added some text about the benefits of the constrained-baseline method. The section now reads (p.11-12):

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).

I also found the end of the first paragraph of page 11 to be a bit confusing. I don't seem to understand why there is a test for the difference between the means of t1 and t2 in each group, as opposed to a test of the difference between the t1 means across groups and the difference between the t2 means across groups. This is what I was expecting but what not what was described. Perhaps this reflects a larger issue with the explanation of the setup of the model or perhaps could require some additional explanation. In the same section, I found myself feeling unclear about how exactly using the blocks accounts for multidimensionality, so perhaps this could be explained more.

You are correct about the comparison of utilities means across groups, not to one another, however, two utilities are tested at once. We have clarified this on p.13:

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 t_1 would be constrained to 0 for model identification, then t_2 and t_3 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.

We further clarified this in discussing the Wald test and hope other modifications we made to our discussion of the second-order model make it more clear (p.6):

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 ($\mu_j - \mu_k$) verse just a utility mean (t_j) 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.

Within the area of psychology that I work in, forced choice and rank order questions are not terribly popular. While I think this paper is still valuable I felt there could be more of an accounting of where these types of items are commonly used, or perhaps even more a pitch for using these types of items in contexts where they are not currently used. Perhaps prior research on these response types has identified important pockets of research communities that specifically use this kind of item and that could be described a little more in-depth.

We have included a description of where these types of tests are used in the introduction (p.3)

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).

I'm curious if the authors have any kind of hypothesis about why the free-baseline approach performed so poorly with large sample sizes in Lee et al (2021). This seems particularly relevant to the current study, as the poor performance was at the sample size selected for the current study. Was that done on purpose? Is there a rationale for why such a method would get worse with large sample sizes?

This was an inaccuracy on our part. The free-baseline method only performed more poorly when impact was introduced. However, the results were consistent across sample sizes when there was no impact. We have clarified this (p.14):
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.

The Present Study

In general, I found the research questions to be a little difficult to understand. The wording of RQ2 is a bit awkward, I think because the question connection between the beginning and the end is lost given the length of the middle clause which summarized the result from Lee et al. 2021).

We agree this question was confusingly phrased. We have rewritten it (p.15):
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?

In RQ3 it's not clear what the question is asking, "more accurate" than what?

We forgot some words. Our apologies. It now reads (p.16):
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?

RQ 5 and 6 are specific to the free-baseline approach, though it's not clear why (as compared to the other methods). It seems like it's likely a foregone conclusion that getting anchors wrong will harm the accuracy of the statistical models. Which I think limits how interesting RQ6 is. It seems like the underlying question, when reading the paper, might be more about whether certain methods are more robust to misspecification of anchor items than others. But that is not really how the question is framed.

We added a line indicating why RQ 5-6 (now 6 & 7) do not apply to the constrained baseline (p.16):

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.

We agree that it seems likely that increasing the number of incorrect anchors will decrease the accuracy of DIF detection, but we would like to examine the severity of the situation. It also seems likely that more anchors is better, however, the purpose of this question is to determine how small of an anchor set size can be used to still accurately identify DIF items. We have added this further context to the interpretations section of appendix 1.

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.

The latter RQs could be more specific that they are looking only at the second-order models and not the first order models.

We added further clarification to indicate we are talking about the second-order model.

Methods

There are still some elements of the methods which do not seem particularly clear. Specifically it's not clear what is meant by saying that "trait correlations will be mixed across all conditions." I initially thought this meant that trait correlations were in some way permuted (mixed) but I think what is meant is that the sign of the correlation is sometimes positive sometimes negative. The term mixed is used for so many things, I'm not sure it's an apt descriptor on its own. Similarly, I found the information about the correlation matrix somewhat contradictory, where in one place it says that it will be matched exactly, and in other places it says it will be randomly generated.

We have rewritten much of this section to make it clearer and also broken it up so the information about negatively keyed items is in its own paragraph. This section now reads (p.18):

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 in 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 means that two traits were negatively correlated with all other traits and positively correlated with each other in 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.

It's unclear from the description of the study design whether the analysis model is a within-factor (i.e., each method is used on each dataset that is analyzed) as compared to a between-factor (data generated is unique to each analysis model). My understanding that the within- method is frequently used to reduce computation time, but either would be acceptable. It should just be clear in the methods.

Our plan is consistent with what you describe as "within-factor", and have specified this at the end of the data generation section (p.23):
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 all generated. Then, they 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.

I do not understand Step 4 of the data generation. Measurement errors should be unique to each person, but the equation given is fixed based on lamba (a parameter that does not vary by person), so it's not clear if this is meant to reflect the variance of the measurement errors. But as written I'm not sure this describes a process that would generate measurement errors.

This was a mistake on our part. As written this was indicating the item errors. Error terms for each person will be generated from a normal distribution. We have corrected this (p.22).

1. Item errors (ε) which were calculated as $1 - \lambda^2$.
2. Measurement errors for each person were generated from a normal distribution, $N(1, \text{sqrt}(\varepsilon))$ for each item response.

Analysis Plan

Using \beta as power us somewhat unintuitive as \beta typically denotes type II errors rates, and so power is typically 1-\beta.

Thank you for the catch. We have made the interpretation more intuitive and clear (p.28):

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.

Reviewer 1 - 17 Oct 2023 09:58

The manuscript describes a planned simulation study investigating uniform differential item functioning (DIF) in Thurstonian item response models (T-IRT). Designed as a replication and extension of Lee et al. (2021), the authors describe several simulation conditions to evaluate the effects of, among others, anchor set size and model misspecification on DIF detection. The manuscript is well written and addresses an important topic for psychological assessments with forced-choice items. Also, the planned simulation study seems reasonable to me and likely will allow to answer the research questions. Therefore, I have a few comments that might help the authors improve their work.

Thank you very much for this positive feedback and your numerous helpful comments. It is truly appreciated!

1) The abstract would be informative if it emphasized that DIF was evaluated for Thurstonian IRT models that overcome the limitations of traditional scoring schemes which result in ipsative scores.

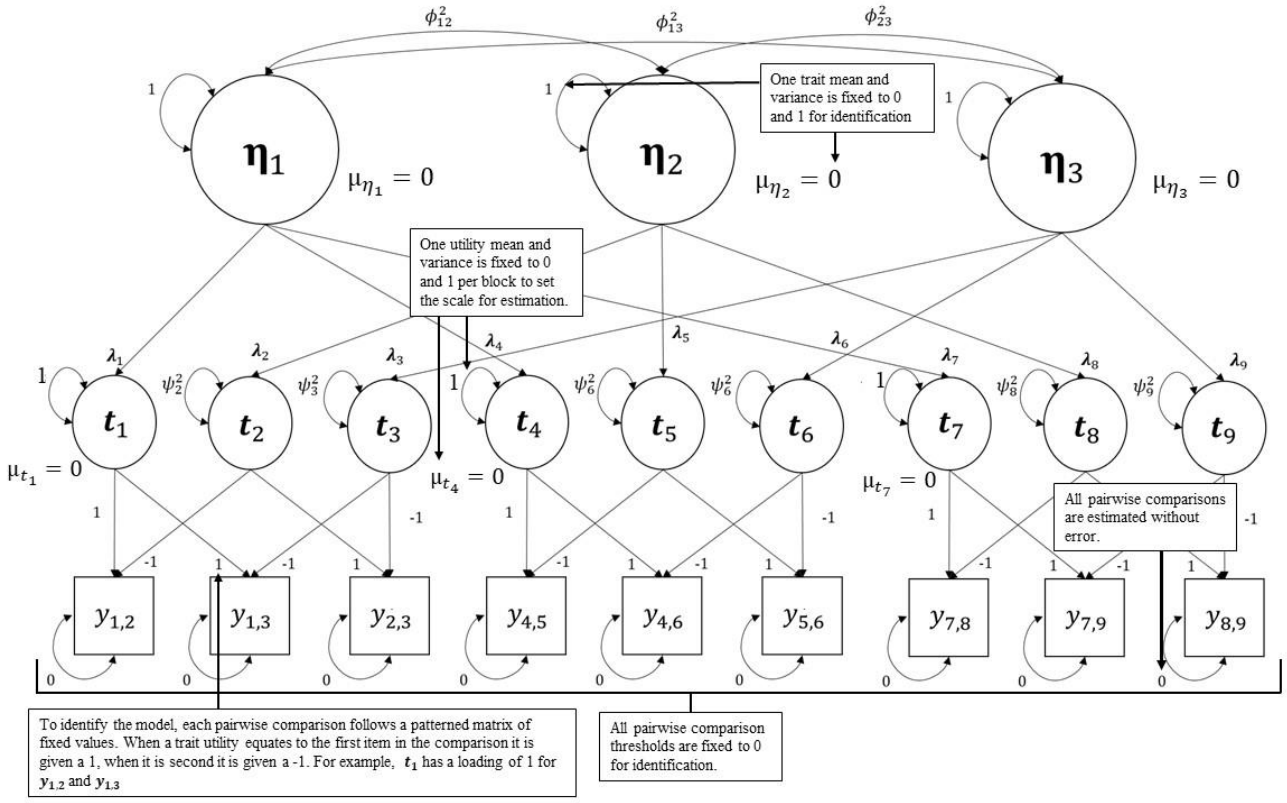
Thank you for this note. We have added some additional context to the abstract to address this issue. The changed areas now read (p.2):
Advances in methodology, such as the usage of the Thurstonian-IRT model to overcome the ipsative nature of FC data, have increased the ease of evaluating the psychometric quality of

forced-choice assessments and spurred an uptake in their use in applied educational, industrial, and psychological settings.

2) The figure presenting the second-order T-IRT is Figure 2 (and not Figure 1). Moreover, I believe the observed scores should be denoted by "y" and not "y*" to correspond to Equation 1. The figure would be more informative if the mean structure would be included as well. It could also be helpful to include the identification constraints described in the text, for example, the fact that 1 uniqueness and 1 utility mean per block are fixed to 1 and 0, respectively.

Thank you for catching this typo. We have changed it to be consistent.

We appreciate your feedback on the figure. We have made edits that are in line with yours and Reviewer 2's feedback. The figure now looks like this:



3) Because the authors plan to use a multiple-constraint Wald test to identify DIF, it might be informative to formally describe the respective test. I was also wondering whether the authors plan to correct the p-values for multiple testing.

We have clarified that this is a multiple-constraint Wald test and provided a brief description of it in the IRT Model for DIF section. It reads (p.14):
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

4) Although the authors describe the necessary identification constraints for the single-group T-IRT, it might be informative to also describe the respective constraints in the multi-group context. I assume in the multi-group models some latent factor means and variances will be freely estimated to acknowledge between-group differences.

Thank you for this note. We added some text in the Model Identification section on how this extends to the multigroup case. This reads (p.7):
When this model is extended to multiple groups in the context of DIF testing, all constraints mentioned extend to both groups. Additionally, the utility error variances are set to be equal 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.

5) The authors plan to use Mplus for their simulation. I was wondering whether the model might also be estimated using an R package (e.g., lavaan). I am not suggesting that this would be preferable in the present situation. I was just curious whether Mplus is currently the only software to estimate these types of models.

The model is implementable in R using the thurstonianIRT package or custom code, however, we found it simpler for our use case to use Mplus.

We have added a note about this (p.23):

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.

6) The authors should adopt a consistent terminology. Currently, “Thurstonian-IRT”, “T-IRT”, and “TIRT” are used interchangeably.

Thank you for the note. We have changed all instances of T-IRT to TIRT. We now also use Thurstonian-IRT until it is fully discussed in the Thurstonian-IRT section and then switch to TIRT.

7) A serious limitation of the planned study is the exclusive focus on uniform DIF. The simulation could be substantially strengthened if non-uniform DIF was evaluated as well (similar to Lee et al., 2021).

While we agree it would be useful to examine non-uniform DIF, we believe this to be outside of the scope of our work and a place for future research as it would double the number of conditions and computation time. We do believe we should justify this decision though and have added text in the Block Factors section. It reads (p.19):

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.

8) The authors plan to simulate two different sizes of DIF that were chosen based on previous simulation studies (see Page 17). I was wondering whether the chosen values are also

representative of applied settings. Are these effect sizes common, for example, in educational assessments or other contexts?

This is a good question. We were not able to locate any additional information such as a meta-analysis on DIF effect sizes in IRT models in applied contexts. We based the effect sizes on the work of Lee and Colleagues (2021) and other simulation studies we cited in text. We have also added this as a limitation of the study (p. 33):

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.

9) The description of the simulation design does not inform about the direction of DIF. Thus, will DIF always favor one specific group or will DIF for different items favor different groups?

Our apologies for the lack of clarity. We have added a clarification in the Block Factors section that reads (p.20):
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 have 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).

10) The simulation design does not specify the sample sizes of the two groups. Are the authors planning to simulate equal sample sizes in the two groups?

We have added two new conditions to the simulation with a larger sample size and an unequal/equal condition and described this on p.19. We also specified the size in each group:
We simulated data such that there were either 1000 or 2000 total responses, just as Lee and colleagues (2021). We also expanded on Lee and colleagues by including an equal and unequal sample size condition. 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).

11) On Page 17, the authors describe how they plan to simulate DIF in the blocks. Because each block will only include 1 item with DIF (see Table 1), it might be helpful if the description emphasized that the authors plan to simulate 10% to 20% of blocks with DIF. Currently, the description focuses on items with DIF which is certainly correct but probably not what the authors want to emphasize. Moreover, the simulation conditions are inconsistently described. On Page 17, the authors report that they plan to simulate 10%, 15%, and 20% of items with DIF. In contrast, Table 2 gives values of 40%, 50%, and 60%.

Our apologies for the lack of clarity. We have added additional context to make the section clear. The main problem was we discussed the number of items with out pointing out how many DIF blocks this resulted in. This also makes where the values in Table 2 come from clear. The section now reads (p. 20):
The effect of the number of DIF blocks on DIF detection was examined by manipulating the percentage of items with DIF. When a block contains an item with DIF, we considered it DIF

block. We tested if there is a difference in the accuracy of DIF detection when either 10, 15, or 20% of the items display DIF. This equated to 40%, 50%, or 60% of blocks with DIF.

12) It was unclear to me how the percentage of DIF blocks (RQ4) and misspecification (RQ6) will be implemented in the simulation. If for example, 10% of DIF blocks are simulated (RQ4) will part of these DIF blocks be included in the anchor block set (RQ6) or will additional DIF blocks be simulated specifically for the anchor block set?

We have further clarified this in the Model Misspecification section on p. 21:
In the free-baseline approach applied to FC data, in this study a model misspecification is when a DIF block is used in the anchor set. We manipulated the amount of misspecification 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 percent 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.

We also further detailed the differences between data generation and analysis conditions to make everything clearer (p.23):

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 all generated. Then, they 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.

Reviewer 2 - 14 Oct 2023 19:27

The current manuscript involves a simulation study that evaluates the impact of model misspecification on DIF detection. The goal of the simulation is to extend the work of Lee et al. (2021) by evaluating the performance of the second-order T-IRT model in detecting DIF under more realistic conditions (e.g., when anchors are unknown).

I thought the authors did a nice job of laying out the motivation for the paper, providing enough background, and highlighting the current gap in the literature in a compelling way. My primary concern is with where the focus is placed in regard to the interpretation of simulation results. In my opinion, the current interpretations are a bit black-and-white, and I think shifting the focus to more nuanced aspects of the results would strengthen the authors' arguments as well as offer more prescriptive guidance to applied researchers. Below, I have included more detailed comments for refocusing interpretations along with more minor comments. The comments below are ordered based on my sense of their importance.

Thank you very much for the positive feedback and these very helpful comments.

1. In my opinion, RQ6 (p. 31) is the most important and most interesting research question. However, Hypothesis B and the first interpretation are hard to follow (for me, anyway). First, Hypothesis B is referring to the size of the anchor set, which seems to be more in line with RQ5 (p.30). Second, for the first interpretation (if there is support for both hypotheses), the second sentence ("However, even in case of complete misspecification…") seems to contradict the first sentence ("If we find support for both…").

Given the importance of RQ6, I think it would be worth giving more consideration to the potential nuances that may emerge from the simulation results and dedicate an appropriate amount of space in the discussion to the interpretation of the results regarding this research question. Currently, the interpretations seem a bit limited. There is one for if both hypotheses are supported, one for if Hypothesis A is supported, and one for if Hypothesis B is supported. For this 72-cell design, there may be some interaction effects worth mentioning.

In regard to one of the potential interpretations for RQ6, the authors state that reducing model misspecification is optimal. I find this interpretation, in and of itself, to be quite uninteresting and one that most methodologists would already agree with. In line with my comments above, I think it would be more compelling to shift the focus of the interpretation to expected interactions or highlighting which specific conditions researchers should be particularly vigilant about when testing for DIF.

Thank you for the note! We added an additional row to Appendix 1 called 'Interactions' that discusses the interpretations of various interactions in line with your feedback. We wanted to focus primarily on interactions with the model specification condition. We have chosen to examine:

a. Anchor set size X Model misspecification.

b. Anchor set size X Blocks with DIF

c. Model misspecification Block with DIF

d. Anchor set size X Model misspecification. X Block with DIF

A. 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.

B. 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.

C. 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.

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.

Hypothesis A and the first interpretation for RQ5 (p. 30) seem to be something that most psychometricians would already agree upon. In cases where the anchor is known to be pure (simulation studies), I think it is safe to say that a larger anchor set will always be better. Given that this set of items is used to estimate the means and SD differences between potential DIF groups, having more DIF-free items should result in better DIF detection. This has been shown to be the case for more traditional IRT models even when an anchor is not pure (e.g., Kopf et al., 2015). Given this, and similar to my comments for RQ6, I think the interpretation for the results of the research question should be refocused (e.g., potential interactions, conditions that are particularly impactful for DIF detection).

Thank you for this reference and your feedback! To your point about RQ 6, we agree that these were confusing. We have rewritten the hypotheses to the following.

A. 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).

B. 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.)

We have also changed the interpretations to be clearer, remove contradictions, and further consider what the results may mean.

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.

In the authors' proposed interpretations for RQ1 (p. 29), they state that if a smaller number of traits results in improved DIF detection, this would support limiting the number of traits measured by an FC assessment.

I would avoid making recommendations of this nature that encourage researchers to add or remove traits from an assessment with the goal of improving DIF detection. I think most psychometricians would agree that adding unnecessary traits or removing necessary traits would negatively impact the quality of a measure. For example, it is known that increasing the length of a test will generally improve reliability (i.e., internal consistency). However, it is not recommended to add items to a test with the goal of (artificially) improving reliability. To avoid spuriously inflating reliability, it is good practice to consult content experts or reference substantive theory, for example. I think the same logic applies to the relationship of trait size and the accuracy of DIF detection. Regardless of the simulation results, "assessments [should] be designed more intentionally only to measure what is needed rather than adding additional factors to examine DIF accurately." This is something the authors state as a potential interpretation on p. 29, but again, I think this should always be the case. If the authors do happen to find an effect of trait size, I suggest presenting it as a consideration researchers should make when evaluating DIF as opposed to it being a motivation for altering trait size.

We agree this was too heavy handed and have changed the interpretation to:

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.

4. In the first interpretation for RQ2 (p. 29), the authors state that improved DIF detection when the DIF effect size is larger would support the use of the latent scoring approach. Although I agree that this provides further evidence for the viability of the approach, I think the study would benefit from including a comparison with at least one other model/approach. As the authors mention in the second interpretation for RQ2, findings in the opposite direction would suggest the need for a different method. Incorporating this in the current study and having results that show how the method the authors currently propose performs in relation to another DIF method would provide a useful frame of reference.

Thank you for this feedback. Our study is focused on the latent scoring approach because it was the only viable method we would find in the literature. As a part of the background for this project we examined Mantel-Haenszel, Delta-Plot, standardization, and logistic regression observed score methods. However, all of these methods assumes unidimensional items, which is not possible with FC items responses. We are not sure if an observed score method could be developed, thus we focus on determining which latent-scoring configuration (Free or Constrained- Baseline) performs best in this study under model misspecification. We described the issues with observed score approaches on p.11.

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.

Minor comments:

5. Sample size is constant in the simulation. Although the authors provide a reasonable justification for this, I think readers will be left to wonder how much sample size could have affected DIF detection. For example, what if doubling the sample size to 2000 counteracts including DIF items in the anchor? If the authors choose not to add another sample size condition to the simulation, I would suggest providing evidence from previous work on the effect of sample size on DIF detection and/or mentioning sample size being kept constant in the limitations.

Thank you for this note. We have added a 2000-person condition to investigate this issue. In line with other feedback, we also introduced an unequal/equal sample size condition. This is described on p.19:

We simulated data such that there were either 1000 or 2000 total responses, just as Lee and colleagues (2021). We also expanded on Lee and colleagues by including an equal and unequal sample size condition. 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).

6. Although Figure 1 and the context that surrounds the use of the term "ipsative" are helpful, it might be worth providing a more explicit definition of "ipsative." This would be especially helpful if the target audience is not likely to be familiar with FC assessments and how they are analyzed.

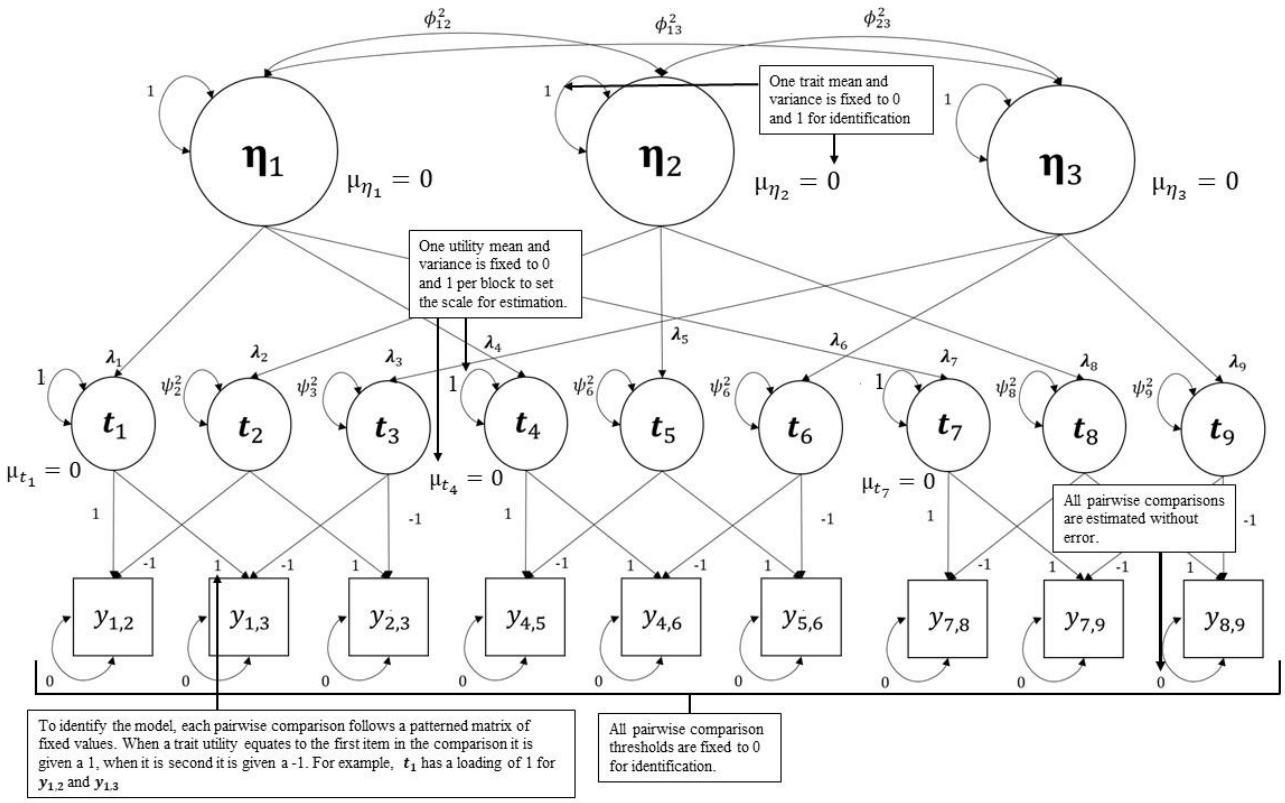
We agree that this would be helpful. we have added a definition to the very start of the manuscript (p.3):
While a well-designed FC assessment can effectively reduce response bias, it produces ipsative data. Ipsative data occurs when the responses are directly dependent on each other: e.g., if you rank "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 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.

7. I think the path diagram labeled "Figure 1" is supposed to be labeled "Figure 2." In addition to the figure title, there is at least one reference to this path diagram in the introduction that will need to be revised accordingly.

Thank you for the catch. We have changed this to be Figure 2.

8. The path diagram labeled "Figure 1," which may actually be "Figure 2," has no caption. Including a detailed caption that allows the figure to stand on its own may be very helpful for readers, especially because it will allow readers to understand the figure without having to go back to the text (some readers may not understand the path diagram without some context).

Thank you for the note. We have added annotations to the Figure to help it stand on its own. It now looks like this:



9. On p. 5, third line from the bottom, the authors refer to "equations 2/3" in a comparison between the unit of analysis for the first-order model and the second-order model. Are Equations 2 and 3 the equations the authors meant to reference?

We have clarified this section. It now reads (p.6):

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 ($\mu_j - \mu_k$) verse just a utility mean (t_j) 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.