Dear Dr. Montoya,

We thank you for your continued support and feedback on our manuscript. We appreciate the opportunity to address these last few points from Reviewer 1. As the feedback is minimal, we will address it directly below. Our responses are in red, with changes to the manuscript highlighted. These changes are also in track changes in the manuscript.

We also wanted to note that we caught an error in our post Round 2 submission. We realized that a constraint for the model could be improved, and our data generation for the errors was incorrect. We initially indicated that the errors should be set equal across groups but determined that setting them to 1 would be best to eliminate their potential impact on examining DIF. We changed this on page 7:

Additionally, the utility error variances are set to be one in both groups, along with a subset of the utility means (*t*).

Figure 2 also reflects this change. We have also corrected this for the data generation in the manuscript on pg.23:

1. Item errors (ε) were set to 1 in line with constraints placed on the model.

We would like to thank you again for your consideration.

Editor:

Thank you for completing your recent revision. Both reviewers indicated that your revisions addressed their concerns very well. There are a few remaining, small comments from Reviewer 1. I believe these can be addressed in a minor revision, and I do not expect to send this out again for peer review.

Thank you very much for your feedback in Round 1 and continued support of the work. We appreciate it.

Reviewer 1

The manuscript describes a planned simulation study investigating uniform differential item functioning (DIF) in Thurstonian item response models (TIRT). I have previously commented on the manuscript and am pleased to read the revised version. I would like to thank the authors for addressing my suggestions. Therefore, I have only a few remaining comments, mainly related to some minor clarifications.

Thank you very much for your feedback. It has made the manuscript and study much stronger. We also thank you for the extra points you offer here.

1) In Figure 2, all latent trait factor means and variances (eta) have identification constraints. Therefore, the comment in the top box may be somewhat misleading because it refers to "one" mean and variance, although all means and variances are constrained. I was wondering whether the latent means in the first-order TIRT (Figure 3) also require zero constraints on the latent means of the latent traits.

Thank you for this catch. We have changed the box to state:

Trait means and variances are fixed to 0 and 1 for identification.

In the case of the first-order model, no constraint is placed on the latent means, and setting the variances to 1 is sufficient for setting the scale origin. We have added a note to the diagram box to indicate this:

Factor variances are set to 1 for identification. Their means are freely estimated.

2) On Page 13, it should correctly read mu_t1, mu_t2, and mu_t3 and not t1, t2, and t3.

Thank you for flagging this error. We have changed them to read correctly.

3) It might be helpful to readers if the introduction (e.g., in the section "Present Study") clarified that the focus of the simulation study is on uniform DIF and that non-uniform DIF is not addressed.

Thank you for the note. We have added a sentence to specify we are focused on uniform DIF in this study on pg.15:

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.

4) I did not understand how the authors derived the percentages for RQ4 on Page 20. If 10% of the items in the 5-trait condition with 20 blocks and 60 items exhibit DIF, then 60 / 10 = 6 DIF items will be simulated. Since each block contains only 1 DIF item (see Table 1), there are 6 DIF blocks. This is therefore 6 / 20 = 30% and not 40% as stated in the manuscript.

Our apologies for this mistake. We made an error in the item percentages. We addressed this by stating the total number of items with DIF instead on pg.20:

We tested if there is 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.

5) In my opinion, simulation studies need to justify their choice of simulation conditions in order to emphasize for which real world scenarios they might be representative of. Therefore, I would hope that the authors could provide some information on the applied settings in which the chosen values of, for example, sample sizes and DIF effects are typical. Simply stating that previous simulations have used these values is not very convincing (Page 20).

We agree this is important to address. We have added some context to various conditions.

Sample Size p.19-20:

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.

Number of Traits p.20:

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).

DIF Effect Size p.20

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.

The analysis conditions (p.21)

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.

Reviewer 2

I have carefully examined the authors' responses to editor/reviewer comments as well as the revised manuscript. Overall, I appreciate the authors' attentiveness and responsiveness, and believe the authors have sufficiently addressed the comments and suggestions raised in the first round of reviews.

In my opinion, the revised manuscript is significantly stronger, and I look forward to the findings of the study. Therefore, I would like to express my support for the acceptance of the revised manuscript.

Thank you for your support and feedback in helping the work reach its present state. We appreciate your time and consideration.