Dear Editor and Reviewers,

Thank you very much for taking the time to read our manuscript and to give thorough feedback to improve the quality of the tutorial. We feel that your exceptionally extensive, detailed, and critical feedback has helped us tremendously to improve the quality of the tutorial.

Below are our responses to your comments and suggestions, and please note that we have copied and pasted the changes we made to the manuscript and the annotated R Markdown file we included as supplementary material just below each comment, to hopefully minimize the need to go through different documents at the same time.

REVIEWER 1

The authors briefly introduce GRMs and launched into a seven-step tutorial with R code using an applied example. They discussed assumption testing, model interpretation, and specification.

I want to first provide context on how I thought about my review. I am a novice to the graded-response model but knowledgeable about IRT methods generally. As I read the tutorial I was focused on if I left feeling like I could now talk about GRMs and apply them. By the end I do think I could confidently say I knew more about GRMs and had the tools to specify a model. I commend the authors for the accessibility of the article, the use of code in open-access software, and crafting a generally well-written manuscript.

I have several points of feedback that I think need addressed before the manuscript is ready for publication. Generally, I think the introduction is weak and that the motivation for the article needs reframing. I think the utility of the article would also be greatly increased if the authors considered when and when not to use a GRM as well as different test formats where it is applicable. Finally, I think the tutorial could be improved by adding a simulated example for an 'optimal' situation that researchers can compare their models against.

I detail these points section by section below.

Introduction

1. I generally think the introduction is far too broad and unneeded as written. I do not believe the authors need to motivate their tutorial by problems with measurement in the field at large. This paper will not fix that. I generally have an issue with tutorials needing a reason to exist other than being useful. Their tutorial is needed because one does not exist that explains GRMs with such a user-friendly focus (if at all from what I found in a quick search). The tutorial is good, has didactic utility, and stands as a great contribution regardless of if it solves some broad issue in the field. I encourage the authors to revise the intro and motivation for the tutorial to make these points.
2. Flake et al., 2017 is miscited in the first sentence of the paper. Their study targeted social and personality measures. It is incorrect to generalize to the entire field of measurement.

Thank you very much for this critical feedback! We realize that our introduction should be stronger and focus on the didactic utility of the tutorial, rather than touching on broader issues that arguably

cannot be solved by creating tutorials. Therefore, we have removed the first part of the introduction entirely and rewritten it as follows (p. 3):

> *"To examine the quality of psychological measures, applied researchers implement statistical or formal mathematical models, which always require specific testable assumptions. Classical test theory (CTT) is one of the model-based measurement theories widely used to examine measurement precision, i.e., how accurate a measure is at capturing underlying psychological constructs and minimizing measurement error. While CTT is useful for informing researchers about measurement precision at the sample level, some want to look closely at specific individuals in their sample or items in their scale. In this sense, item response theory (IRT) fills this need by allowing researchers to inspect their measurement quality at the person and item level. IRT models are nonetheless known to be computationally intensive and less practical than CTT-based analysis, so a tutorial with didactic value can potentially help applied researchers to consider IRT analysis as a tool for their research.*
>
> *The purpose of this tutorial is to provide a brief introduction to the graded response model (GRM), a family of IRT models specifically designed to assess the measurement precision of a polytomous (Likert-style) scale. In this paper, we aim to keep our tutorial concise, so we only briefly introduce the basics of IRT concepts. Interested readers are referred to available didactic texts on IRT, such as Embretson and Reise (2000) for the introductory level, and Baker and Seock-Ho (2017) and De Ayala (2022) for a more technical and comprehensive overview."*

3. Please define measurement precision the first time you mention it.

Great point! We define the measurement precision in the introduction as follows (p. 3):

> *"…Classical test theory (CTT) is one of the model-based measurement theories widely used to examine measurement precision, i.e., how accurate a measure is at capturing underlying psychological constructs and minimizing measurement error."*

Graded Response Model

1. I think this section is well-written and clear. The graphs are also very good.

Thank you for your positive feedback!

2. Pg.5 - "While both items might measure the same latent trait (i.e., trust in science), those who trust scientists working in natural science are very likely to also trust biologists."

   We expect some level of correlation between items on the same trait and I think this example implies correlation may constitute a violation of local independence. I think it could be made clearer what a true violation of the assumption is/looks like. The authors do caution against throwing out these items when checking the pairwise correlations, but it is an important point to explain here too.

We totally share the point. We add more context to provide a clear description of how local dependency may look like when fitting a model, as follows (p. 13):

"Note that evidence of unidimensionality does not always warrant local independence. As an eyeball example - consider a scale measuring trust in science that includes: "I trust scientists working in natural science" and "Most biologists are trustworthy." While both items may measure the same latent trait (i.e., trust in science), as indicated by moderate or large loading factors, those who trust scientists in the natural sciences are very likely to trust biologists as well. Therefore, in practice, it is likely that these items will still be moderately correlated even after taking into account their loading factors on θ."

3. I'll put this point here but it applies to the entire paper. I would like the authors to also consider or at least describe different test formats where GRMs are useful. Situational Judgement Tests come to mind, but I think this tutorial will have more utility if the various test formats a GRM may be useful in are described. I do not think you need an R tutorial on each, but they should at least be discussed in text.

Good point! We add this as you suggested (p. 9):

"GRM generally can be applied to a different types of test formats, such as scales measuring behavioral outcomes (Arias et al., 2016), personality and attitude scales (Rauthmann, 2013), or patient-reported outcome measures (Normand et al., 2006). Situational judgment tests (SJTs), which measure non-cognitive abilities, decision-making, and problem solving in context-specific situations (Corstjens et al., 2017), may also be suitable for GRM analysis when the response categories are ordered."

Illustrative Example

1. I appreciate the use of a real scale with real data, but I do think it would be useful to include an optimal example using simulated data as part of the tutorial. This will provide readers with something to compare against.

We agree that this is a good idea. We have added an additional part in our annotated R Markdown file and have mentioned this in the manuscript, as follows (p. 15 and 16):

"*We also include a longer, annotated R Markdown file (.Rmd) for the example we used as a supplementary document, which we highly recommend using for a didactic purpose. In the same file, we also include a part where we demonstrate the application of GRM using a simulated dataset, which we do not include in the article, so that readers can compare the example we present here against an ideal scenario." (p.15)*

"The dataset provided in our repository has been cleaned, with all unfavorable items reversed scored, and is ready for analysis. Since the dataset is openly available on the Open Psychometrics website, readers can also import the dataset directly from the website into their R environment. For interested readers, we show in detail how to perform the import and data cleaning process in our annotated R markdown file, which is available as supplementary material." (p.16)

Below is the snippet of the second part of the annotated R markdown file, which is also available as supplementary materials.

"THE SECOND PART: Simulated Dataset

To give readers a clearer picture of the implementation of GRM, in the second part, we will demonstrate how to simulate a dataset using simdata() function in mirt package. We will simulate a new dataset using model and item parameters that we estimated with the real data in the first part so that readers can compare the first (real) and the second (ideal) model.

Step 1: Simulate the dataset using mirt package

Simulating a data set should always start with specifying the parameters, and to that end, we will replicate a dataset using model and item parameters that we have estimated in the first part. mirt package also provides the option to simulate response patterns (i.e., scale data) from a custom input matrix of theta, but we don't include the example here in the material. Those who are interested in simulating response patterns from a custom theta matrix are advised to read the examples provided in mirt documentation.

It's important to remember that the simulated data is constructed by strictly adhering to the model assumptions and parameters. Therefore, model and item parameters here are just statistical artifacts that should not be taken at face value. The purpose of implementing GRM with simulated data here is to show how the "ideal" scenario looks like.

To simulate response patterns from estimated RWA model we have shown in the first part, type the codes below.

{r}

sim_data <- simdata(model=fit, N=800) # simulating 800 response patterns from RWA scale we previously estimated in the first part

sim_data <- as.data.frame(sim_data) # and then set the data to a data frame

We now have a newly created data frame, sim_data, containing 800 response patterns estimated from the model and item parameters we already had."

...the rest shows similar steps as modeling the real data, as we have shown in the manuscript.

2. There are data management things in this tutorial that I think just muddy the water. For example, a researcher specifying their model would not need to uncover what the 0's mean or which items need to be reverse coded. They would know as they design the assessment. I think Step 1 would be stronger by excluding these data management tasks and providing a clean file, so the focus remains on specifying and interpreting a GRM.

We share this concern, and in fact, removing the data preparation and management parts free up space to add and address reviewer feedback, making the tutorial stronger and clearer for readers. We have removed these parts entirely but left a note for readers to see the process in detail in our annotated R Markdown file (p. 16).

"The dataset provided in our repository has been cleaned, with all unfavorable items reversed scored, and is ready for analysis. Since the dataset is openly available on the Open

Psychometrics website, readers can also import the dataset directly from the website into their R environment. For interested readers, we show in detail how to perform the import and data cleaning process in our annotated R Markdown file, which is available as supplementary material."

3. The Github link does not link to anything.

We don't include the Github link to allow blind review, and added the information on p. 16:

"The Quarto file (and its corresponding .docx and .pdf output) and an annotated R Markdown file are publicly available on a Github repository (unlinked for blind review)."

..but we will add the Github link after the manuscript is (hopefully) accepted.

4. Did the authors make ggmirt? I would make that another selling point of this paper if so!

We wish we had the idea to develop ggmirt! We have properly cited the ggmirt package to give credit to the developer.

5. ds <- read.csv ("data/data.csv"). I think it is better to use a placeholder text as users may need to access the folder from somewhere else.

Great point, thanks! We have changed this as you suggested (p. 16):

"The cleaned dataset (data.csv) and code book (codebook.txt) are available in data folder in our repository. Readers can download them from our repository and then simply import the dataset with the following command:"

```
data <- read.csv("path_to_data/data.csv") # Replace "path_to_data" with th

e actual path to your data file
```

6. Page 6: Please specify the mid-point of 5 explicitly. This will make the -4/+4 make more sense.

Makes perfect sense! We made this change as you suggested (p. 14):

"...which consists of 22 items in which participants are asked to indicate their agreement with the items on a nine-point scale ranging from "strongly disagree" (-4) to "strongly agree" (+4), with 0 as the midpoint."

7. Table 1: Please make it so the minimum and maximum are on one line.

We decided to remove the data preparation and management parts so that Table 1 is no longer present in the manuscript.

8. Pg.17: “The third line of code is a function to calculate a factor loading ( ) and commonality (h 2 λ) of each item, which are also provided in Table 2.”

   This seems redundant with Step 3. I know the authors have it here as it is just a part of the MIRT output, but I think referencing loadings and discriminations within the same section is a bit confusing. I would remove this from Step 4 in the paper and perhaps just leave it as a note in the supplementary materials.

Thank you for your feedback! However, we decide to keep this, since readers who are familiar with modelling in R would naturally use summary() function after fitting the model. Readers would be more likely to be confused if we did not provide the explanations why FA of the data after fitting a mirt object (showed after calling summary() function) slightly differs from EFA we presented in the earlier part. We have provided the explanation why the FA results (loadings and commonality) slightly differ from the previous part as follows (p. 23):

> “*The results are slightly different from an EFA analysis we ran earlier, because mirt runs EFA using a quasi-polychoric correlation matrix, while the one we ran earlier to test unidimensionality used a Spearman correlation matrix as an input. However, most importantly, we see that all items are significantly loaded to one factor, and the factor now substantially accounts for 65.1% of the variance in the data, which strengthens our assumption that the RWA scale is unidimensional.*”

9. Table 2 rocks!

Thank you!

10. I may have missed this somewhere in the tutorial, but I was curious as to why there are only 8 discrimination parameters? I assume this is because the mid-point does not have one estimated. I would clarify that again with Table 2.

Since there is only one discrimination parameter per item, we suspect that you mean threshold parameters? Since the scale has 9 response categories, it makes sense that there are only 8 threshold parameters, because each threshold represents the point of transition between adjacent categories, not the number of categories themselves. For example, *b*1 is a threshold from moving from -4 to -3, *b*2 is to move from -3 to -2, *b*3 is to move from -2 to -1, *b*4 is to move from -1 to 0, *b*5 is to move from 0 to +1, *b*6 is to move from +1 to +2, *b*7 is to move from +2 to +3, and finally, *b*8 is to move from +3 to +4.

To clarify this, we add more explanation as a note in Table 2 (p.25) as follows:

> “*There are 8 threshold parameters because each threshold represents the point of transition between adjacent categories, not the number of categories themselves.*”

11. χ 2 {apatb-table4} – Please go through and QC to make sure the things you are referencing were inserted properly.

Thank you for pointing this out! Apparently, there was a formatting error with the previous version of our quarto manuscript, but we have corrected this and make sure that the errors do not appear in the current revision.

12. For Figure 3, I think it would be more helpful to have just one of the IPFs at a close up.

We present the IPFs of one item at a close up in the Introduction (p. 11), so we think it would be more helpful for readers to skim through what all the IPFs look like after fitting the model, and to get an idea of whether the IPFs are consistent with the threshold parameters estimated in the earlier part. The goal here is to allow participants to see the output when they call a function to create IPFs for all items in the model. Presenting IPFs from only one item would potentially raise more questions from readers as to why only IPFs from one item are shown and not the others.

Discussion

1. I think it would be great if the authors made a flowchart for the process.

We truly like this idea! We made an annotated flowchart detailing the process, and copy and paste the discussion part in your #3 point.

2. Pg.33: "Since IRT is less popular than CTT, we also briefly explain how IRT differs from CTT in its assumptions.."

   I do not know if that is true. Perhaps in social psychological assessments. However, in other areas such as I/O and educational psychology, IRT is used frequently. I think this just comes back to the point of the authors being far too broad in their discussion and intro sections.

We totally share the concern, and therefore we have removed this part.

3. The conclusion is quite short. I would like to see a discussion on when and why you would use GRMs over other methods, where it is applicable, and an overall more in-depth wrap-up on its utility.

We understand your point. To address this, we have added more wrap-ups and the last part of the discussion focuses on strategies to overcome local dependency (pp. 39-43), but we want to keep the discussion concise since the journal has a strict and rather short word limit.

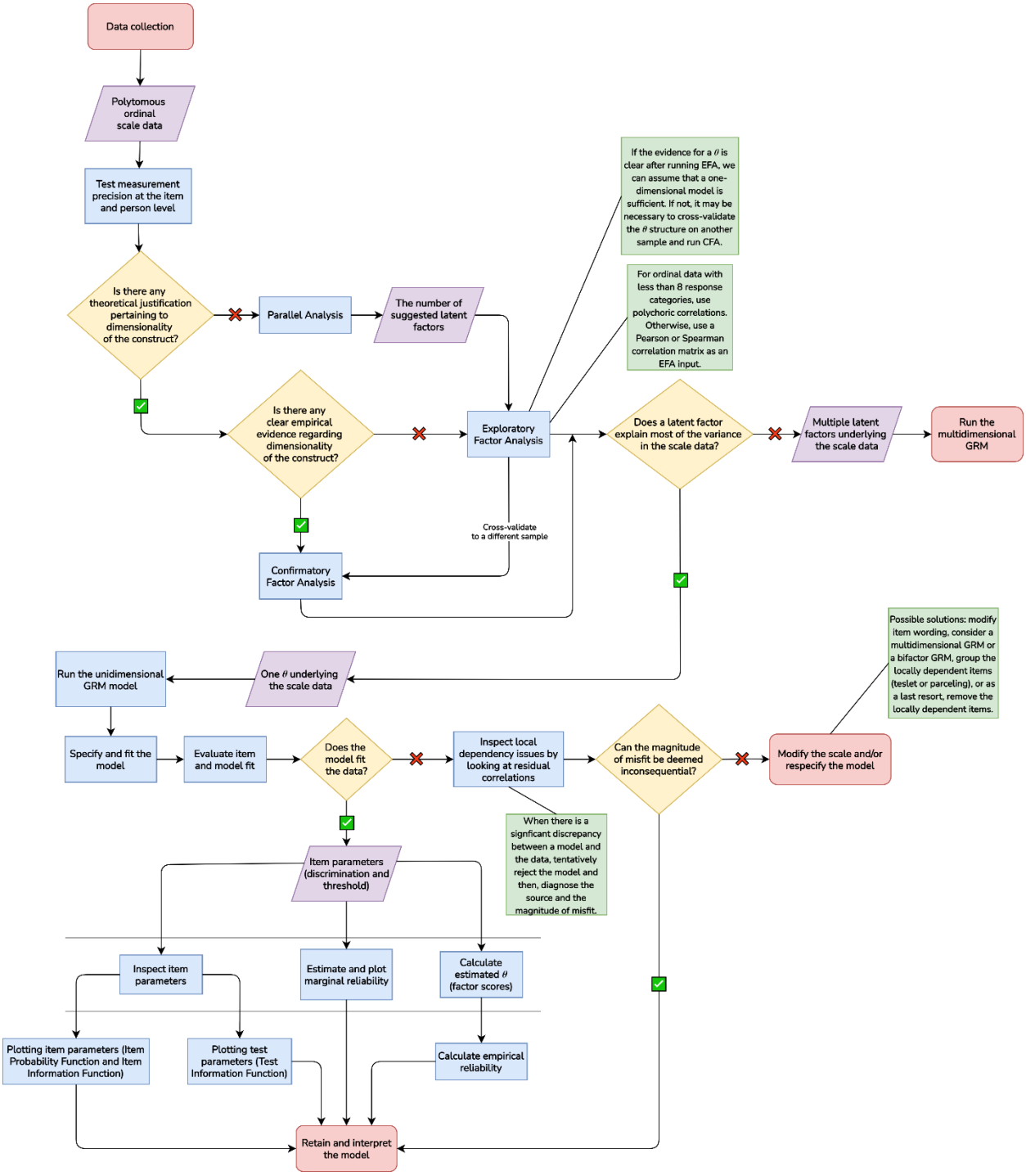
> "*In this tutorial, we have demonstrated the applicability of GRM analysis as part of the IRT family to help applied psychology researchers assess their measurement quality. We provide a non-technical guide to implementing a GRM analysis through a simple 6-step process using a real, openly available dataset. To maximize the effectiveness of this tutorial, we show how to perform a GRM analysis using R, an open-source statistical software, and make all materials publicly available.*

We begin the tutorial with a theoretical overview of IRT, which underlies GRM, so that readers can relate the practical steps to the theory behind the analysis, including an explanation of how IRT differs from CTT in its assumptions. In general, GRM offers several benefits and is useful as a complement to CTT analysis, especially when researchers focus on closely examining the precision of their measures at the item and person levels. However, we emphasize that our goal here is not to argue for the superiority of IRT over CTT, as the choice of analytic tool depends largely on the specific research question at hand. It is important to emphasize that researchers should be aware of the merits and limitations of their chosen methods. Therefore, researchers should always justify why they choose a particular method over many available alternatives.

We have described the process of conducting a GRM analysis, starting with testing dimensionality, fitting the model, calculating and plotting item parameters, and calculating scale reliability. To help readers intuitively understand the process, we have summarized the steps we demonstrate in this tutorial in an annotated flowchart (Figure 7).

Figure 7

An Annotated Flowchart Depicting the Process of Fitting A GRM Model



Note that the model we estimate does not fit the data perfectly due to local dependency, as we showed in steps #3 and #4. If researchers encounter this problem when evaluating their scale data, we recommend tentatively rejecting the model and then locating the source of the model misfit before interpreting measurement precision or drawing substantive conclusions based on these estimates (Kline, 2023). In doing so, researchers can take a closer look at residual correlations to identify the source and magnitude of misfit (Kline, 2023; Maydeu-Olivares, 2015).

Once the sources of misfit are identified, one solution to address them is to rephrase or combine the content of the problematic, locally dependent items and then cross-validate the modified scale on another sample. Alternatively, researchers may need to reconsider the θ structure. If the data are unidimensional, but some of the items are locally dependent, this may indicate the existence of multiple θ within the data structure. In this sense, modeling the test data as a correlated multidimensional model or a bifactor model (i.e., multidimensional models with a g factor) may be a viable solution. Another solution to consider is to group locally dependent items by modeling them together as a "testlet" (Cook et al., 1999), or to combine locally dependent items into a single composite score (i.e., item parceling), especially when the goal of the analysis is to understand the construct being measured rather than to identify the relationship between items and the θ (Little et al., 2013). Removing locally dependent items should be a last resort, as it may improve model fit, but can jeopardize measurement precision.

Finally, while it is important to acknowledge that while a poorly fitted model may result in biased parameters, in reality, no IRT model can be expected to fit the data perfectly (Maydeu-Olivares, 2015). That said, some models may be useful even if they are wrong to some degree (Box, 1976). Therefore, researchers are encouraged to not only evaluate the overall fit of their model, but also perform a piece-wise evaluation in some parts of their model (Maydeu-Olivares, 2015) to get a comprehensive picture of the performance of their measures."

Minor Points:

This is a personal preference thing, but I believe the title should remove the word 'gentle'.

Corrected, thank you!

REVIEWER 2

The manuscript introduces readers to the graded response model (GRM), including how to conduct the analysis in R. The manuscript uses a publicly available data example to illustrate the analysis with topic-relevant data. Overall, the writing is clear and straightforward. However, there are some areas which readers could benefit from increased description and education to ensure those not familiar with the analysis understand the points being made. Recommendations are provided below. I do my best to highlight concepts and term which more information about for the reader would be helpful; however, reading it from that perspective before resubmission may be helpful.

First general recommendation is to provide a bit more context for readers about why and when to use GRM. Putting it in reference to CFA approach as well as an IRT approach they may be more familiar with (such as correct/incorrect with test items) will help researchers then understand when

they'd want to use the GRM approach instead and the type of information they gain to answer research questions. There is work already out there connecting CFA and IRT for the same measures, so this can help provide scaffolding for your readers. These are focusing on IRT with respect to a grouping variable, such as gender, primarily; however, can help situate GRM within the larger measurement quality analysis context.

Reise, S. P., Widaman, K. F., & Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychological bulletin, 114(3), 552.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. Structural Equation Modeling, 18(2), 212-228.

Buchholz, J. & Hartig, J. (2020). Measurement invariance testing in questionnaires: A comparison of three

Multigroup-CFA and I RT-based approaches. Psychological Test and Assessment Modeling, 62(1), 29-53.

Thank you for your exceptionally detailed and critical feedback. We initially included a section giving a brief overview of IRT before going on to explain GRM, which we had to discard because the journal has a fairly short (6,000 words) and strict word limit. We fully share the concern that an introduction to IRT and how its assumptions differ from CTT is important, especially for readers without a background in IRT. We reached out to the editor and asked for his advice - and we are very grateful that he allows some discretion on the word limit, enabling us to put a section entitled “A Brief Overview of Item Response Theory” (pp. 4-8) back in the manuscript, as follows:

> *“To scrutinize their measurement quality, psychological researchers have relied on two theoretical frameworks underlying psychological testing, CTT and IRT, to decompose observed or raw scores into their deterministic (i.e., “true” score) and random (i.e., measurement error) components (Zumbo, 2006). In practice, researchers implementing CTT principles apply structural equation modeling (SEM), exploratory factor analysis (EFA), or confirmatory factor analysis (CFA). Specifying the measurement model of an SEM model, or fitting EFA and CFA models, allows the covariance matrix of a given dataset to be decomposed into latent constructs (θ) representing true scores or actual skill or trait levels, with measurement error (i.e., residuals or unique variances).*
>
> *In this sense, CTT generally assumes that the relationship between observed score and true score is linear, as true score is essentially a linear transformation of observed score (Embretson & Reise, 2000). However, IRT models this relationship differently in a probabilistic rather than a linear fashion. More sharply, an IRT model assumes a probabilistic relationship between the observed score and the latent trait being measured (θ). The probabilistic relationship in an IRT model accounts for the statistical properties of the items, such as item discrimination, difficulty, and, if necessary, pseudo-guessing parameter (Baker & Seock-Ho, 2017). IRT methods were originally developed to evaluate traditional aptitude and achievement tests that measure knowledge and skill levels, but the application of IRT has been extended to personality, attitude, and other psychological scales of various formats.*

The most critical difference between CTT and IRT is the way each principle conceptualizes measurement error. CTT uses standard deviation of the observed score (s) to calculate standard error of measurement. Because measurement error is based on the observed score, CTT assumes that measurement error is sample-dependent and constant for all individuals in the sample, regardless of their θ level. This shortcoming limits the utility of CTT, especially since psychologists often need to interpret individual scores, not just the test or the sample as a whole. In addition, because the standard error of measurement is also a function of reliability, it is assumed that scale reliability will always be the same across different levels of θ. This is unrealistic because researchers often encounter situations in which a scale is "too hard" to endorse for a group of "low-performing" (i.e., low θ) participants, providing little information beyond an indication that their θ level is much lower than what the scale can measure.

IRT offers a solution to this issue by enabling the calculation of standard errors for each individual (SEθ), and test reliability can be inferred from the average of SEθ across a sample of participants (Embretson & Reise, 2000; Lang & Tay, 2021). This approach allows for the estimation of reliability at varying levels of θ (i.e., test information function - TIF). By doing this, when a test is overly challenging for low-performing individuals or too easy for high-performing individuals, IRT analysis can help identify the levels of θ where the test is most reliable. A concrete example of this is a study scrutinizing the reliability of the Short Dark Tetrad (SD4) scale from an IRT perspective (Blötner & Beisemann, 2022). According to this study, the sadism subscale of the SD4 scale is the most reliable for measuring individuals with average to high levels ($x^{-} < \theta < 2.5$ SD) of sadism but suboptimal for measuring those with low levels of sadism (Blötner & Beisemann, 2022). Therefore, IRT models offer a more informative approach, thus are useful for researchers who wish to closely examine the performance of their measures at the person level.

Further, IRT allows researchers to examine the performance of a specific item by specifying the relationship between item score and θ given one-, two-, or three-parameter. One-parameter logistic (1PL) or Rasch model accounts only item difficulty (b) while presuming equal item discrimination (a) across items. Practically, this model slightly overlaps with CTT in the sense that they assume that all items carry equal informative value thus the estimated θ of a 1PL/Rasch model is identical to the sum score of CTT (Lang & Tay, 2021; Stemler & Naples, 2021). In this sense, Rasch model assumes that the latent trait (θ) should remain unaffected by specific items used in the test. For example, the difference in depression levels between two individuals should be always the same regardless of the scale used to measure their depression levels, be it Beck Depression Inventory (BDI) or Patient Health Questionnaire-9 (PHQ-9). While this principle reasonably enforces objectivity in measurement practices, it demands a strong theory delineating the construct and strict requirements of data-model fit, while both assumptions rarely hold in a real-world setting.

Furthermore, the two-parameter logistic (2PL) model fills this gap by allowing item ability to differentiate individuals with varying levels of θ (i.e., item discrimination parameter - a) to differ. Additionally, in some contexts, researchers may suspect that a part of the probabilistic relationship between observed score and θ is explained by guessing thus three-parameter logistic (3PL) model incorporates pseudo-guessing parameter (c).

Further, IRT paradigm has rapidly developed to include various models suited to specific contexts, such as handling ordinal responses (Muraki, 1992; Samejima, 1997), categorical responses (Thissen et al., 2013), or assessing multidimensional traits simultaneously (Bock & Aitkin, 1981; Chalmers, 2012). It is important to note that 1PL, 2PL, and 3PL models are only applicable to binary or dichotomous data (e.g., true/false response), and the focus of this article is nonetheless to show the utility of IRT for fitting the ordered (Likert-style) responses. We briefly summarize the distinctions between 1PL, 2PL, 3PL, and GRM model in Table 1."

Table 1

Comparison Between Common IRT Models

Model	*Key Characteristics*	*Data Type*	*Response Options*
1-PL Model (Rasch Model)	*1. Estimates only item difficulties (b).* *2. Assumes that all items have equal discrimination parameters (a).* *3. Item and person parameters are independent.*	*Dichotomous*	*Correct/Incorrect (0/1)*
2-PL Model	*1. Estimates item difficulties (b) and item discrimination (a).* *2. Less stringent than 1-PL model since it allows item discrimination parameters (a) to vary.*	*Dichotomous*	*Correct/Incorrect (0/1)*
3-PL Model	*1. Estimates item difficulties (b), discrimination (a), and pseudo-guessing parameter (c).* *2. Appropriate for modeling a test data with multiple responses (e.g., multiple-choice tests), and thus, guessing might influence participants' responses.*	*Dichotomous*	*Correct/Incorrect (0/1)*
Graded Response Model (GRM)	*1. Appropriate for modeling ordinal data with more than two response categories (i.e., Likert-style).* *2. Estimates a discrimination parameter (a) and multiple threshold parameters (b) per item.*	*Polytomous*	*Ordered Categories*

Second general recommendation is that individuals be able to read, understand, and follow along without running the R example while reading. This may mean either adding more information to the article itself and/or moving some material into a supplemental file. While there are pros and cons to either approach, it felt that perhaps in trying to keep it all in one manuscript, explanation at time from either the conceptual or the Rscript side was missing due to this difficult balance.

We see your point. We agree that some parts of the manuscript can safely be moved to the Supplementary Material in order to free up some space to explain the substantive concepts. Therefore, we have decided to move the steps for data management and cleaning so that the tutorial begins with the dimensionality test. We explain the change we made by responding to each of your recommendations below.

The rest of the recommendations are in order of appearance in the manuscript.

Page 2

1. Line 32: θ is not defined as was done with b and a.

Thank you for pointing this out! We have corrected this as you suggested (p. 3):

> "*Specifying the measurement model of an SEM model, or fitting EFA and CFA models, allows the covariance matrix of a given dataset to be decomposed into latent constructs (θ) representing true scores or actual skill or trait levels, with measurement error (i.e., residuals or unique variances).*"

2. Line 54: the response categories given are 1 to 4; however on page 3, line 12 the response categories presented are from 1 to 5.

Corrected as below (p. 9):

> "*To illustrate the step function, consider a scale measuring sadistic personality (e.g., "watching a fist-fight excites me," etc.) with response categories ranging from 1 (strongly disagree) to 5 (strongly agree).*"

Page 3

1. Lines 27-31: mention of the mean is brought up for the first time. Introducing how the mean is part of this process before going into the interpretation would be helpful. Expanding this introductory information more for those readers unfamiliar with IRT approaches or IRT with non-binary data will be helpful.

Great suggestion! Added as you suggested (p. 10):

> "*Therefore, a GRM model of a five-point Likert scale calculates four item threshold parameters (b): the location of θ level where individuals are equally likely to respond 1 or 2 (b1), 2 or 3 (b2), 3 or 4 (b3), and 4 or 5 (b4). Each threshold is exactly the point at which a participant is equally likely to respond to either of the adjacent response categories. Note that the level of the latent trait (θ) that a person has is typically centered around a mean of zero. The mean represents the average level of the trait across all participants in the sample. When we interpret item threshold*

parameters, such as b1, we are considering how far an individual's θ level deviates from this mean. For example, if b1 = -1.23 for the item "watching a fist-fight excites me", this means that participants with a sadism level 1.23 below the mean might change their response from strongly disagree to moderately disagree."

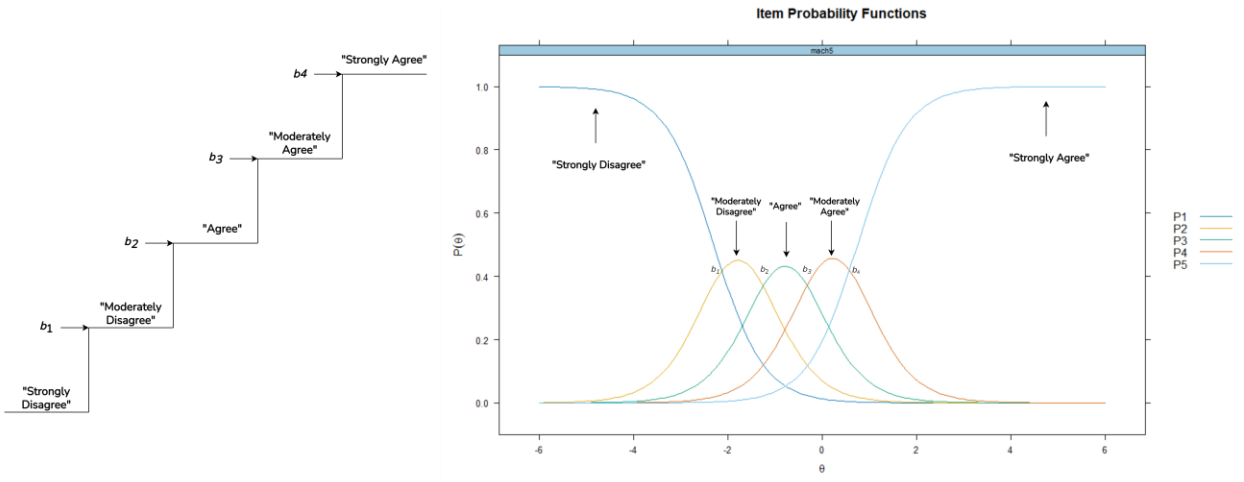
2. Recommend explaining more about the probability curves and particularly the equal probability of endorsing adjacent categories more. Could also anticipate readers being confused why the strongly disagree and strongly agree probability curves look different from the other probability curves.

We have added a few sentences to make the explanations clearer and added a staircase illustration to complement the IPFs, hoping that it is now clear enough for the readers (pp. 10-11).

"Figure 1 shows an example of an item probability function (IPF) from an item on a scale with five response categories along with the staircase illustration. In general, IPF describes a relationship between θ (x-axis) and the probability of endorsing a response category (y-axis - Pθ). Figure 1 consists of five category probability curves, each of which represents the probability of endorsing a response category given θ. Item threshold is exactly the location of θ where two adjacent category probability curves cross each other. The "Strongly Disagree" and "Strongly Agree" curves are boundary categories in a GRM model, which results in their peaks being at the extreme ends of the θ. These curves look different than the other because they represent the cumulative probability of selecting the most extreme responses, while intermediate categories peak in the middle and have more defined, bell-shaped curves. The probability curves of boundary categories simply mean that individuals at extreme θ levels are very likely to select extreme responses."

Figure 1

Staircase Illustration and Item Probability Functions of an Item in a GRM Model



Page 5-6

1. The assumption of local independence of items with a multi-item measure of a construct is unlikely to be met as highlighted by your example. However, it is also a very important point influencing how the analyses are conducted. More explanation regarding the points made in this paragraph will help readers understand the point being made. For example, explaining why looking at relationships with residual correlations rather than just item correlations to further help illustrate this point conceptually to tease out the different parts of the variance being referred to here.

Thanks for the suggestions! We have added more explanation about what readers should expect regarding local dependency issues (i.e., that it is difficult to be met in multi-item scale) and how it looks in practice (i.e., that item pairs are correlated with each other even after accounting for their correlation with θ (pp 12-13). To avoid redundancy, we explain in detail why local dependency is important to assess and how to handle the issues in "Step 4: Model Residuals".

> "*Relatedly, a GRM model assumes that the item pool is locally independent, as the third assumption. Local independence suggests that participants' responses to one item do not affect their responses to another. To test whether items are locally independent, researchers can examine the relationships between item responses after accounting for θ by performing residual correlation analysis (Chen & Thissen, 1997). In practice, it is quite difficult to fully satisfy the local dependency assumptions when researchers measure a construct with multi-item measures, so we should always expect some degree of local dependency to plague the model. That said, some local dependency problems may be insignificant, and there are several strategies to avoid severe model misfit due to local dependency, which we will discuss later.*
>
> *Note that evidence of unidimensionality does not always warrant local independence. As an eyeball example - consider a scale measuring trust in science that includes: "I trust scientists working in natural science" and "Most biologists are trustworthy." While both items may measure the same latent trait (i.e., trust in science), as indicated by moderate or large loading factors, those who trust scientists in the natural sciences are very likely to trust biologists as well. Therefore, in practice, it is likely that these items will still be moderately correlated even after taking into account their loading factors on θ.*"

2. Page 6: given the new terms and symbols readers may be absorbing during this manuscript, it is recommended RWA be written out to reduce unnecessary additional acronyms.

Since we have rather limited space due to the word limit policy, we still use the acronym to save space. However, to avoid confusion, we agree with your suggestion below to use "RWA" strictly to refer to the "right-wing authoritarianism" construct, not the data set or theta in the model specification.

Page 7

1. Line 2, do not need parentheses around Altemeyer, just around the year.

Corrected, thanks!

Page 9-11

1. Page 9-10, Lines 53-32: not a minimum value of 8; providing a value as an example would help readers who are not simultaneously running the example follow with the points being made. It starts to sound from context that 8 represents missing; however, given the use of 8 related to the actual GRM analysis this increases confusion. I recommend making this interpretation of this theta clearer at the start of this discussion or a footnote for people running the analysis, because of the confusion it presents.
2. This is also an example of where such material and the initial reading in and preparation of the data in R could be solely provided in a supplemental file so the manuscript can focus on that which is specific to the GRM, such as the local independence test.
3. Similar point regarding reverse-coding all of this data prep and reading in of packages could be in a supplemental file. I really appreciate the detail and not assuming readers comfort with R, however this is 3 pages that could in the manuscript focus on material unique to GRM conceptually and practically that all readers need, while the detail of these preparation steps can be in the supplemental file to be available to those who will benefit from it. Can reference these steps were done and then start with Step 3 section in detail

Page 11-12:

1. End of page 11 states Table 3 has "mean scores across items does not drastically vary .... "
2. Table 1 is the table on the next page not Table 3
3. Table1 does have mean score values that vary just as described in the original items, pre-reverse coding. This trend seems to continue throughout the data table examples, with the tables presenting the original, raw values rather than the reverse coded values.

A general response to the feedback on pp. 9-11: We have decided to remove these parts completely from the manuscript and put them in the annotated R Markdown file that we include as supplementary material. We also mention this in the manuscript, so that readers who are interested in looking at these processes in detail can look at the R Markdown file (p. 16):

> *"The dataset provided in our repository has been cleaned, with all unfavorable items reversed scored, and is ready for analysis. Since the dataset is openly available on the Open Psychometrics website, readers can also import the dataset directly from the website into their R environment. For interested readers, we show in detail how to perform the import and data cleaning process in our annotated R Markdown file, which is available as supplementary material."*

Page 13

1. The second paragraph quickly goes over the conclusion and why using Pearson correlation matrix as input rather than going with the ordinal approach. While I agree with all these conclusions and decisions, explaining and justifying them for those less familiar with the reasons behind these decisions is important. What if you'd had 5 response categories, would your decision have been the same? This will help future researchers implement and support their implementation decisions with this analysis.

2. Also recommend setting this up with being more clear that rather than using the irt.fa for these reasons … you are going to do the correlation approach.

Great point! We have made this change in the manuscript (pp. 16-17):

> *"`irt.fa()` function is to run an EFA with a polychoric correlation matrix as an input when the data are ordinal, which is often the case in practice, especially when researchers are working with a five-option Likert scale. Since we have more than eight response categories in our current dataset, after running `irt.fa()` function, an error message, "polychoric is probably not needed", should appear in the console. Therefore, another option would be running an EFA manually with a correlation matrix as an input. To determine which correlation method should be used when constructing the correlation matrix, we need to check whether items are normally distributed by calling this command:*

```
skim(data)
```

> *After running this command, readers can see in the console a table summarizing the data frame, including histograms of each item in `hist` column. Here, readers would find that all items are heavily skewed, so that we should use a Spearman correlation matrix as an input for EFA. In doing so, we need to create a Spearman correlation matrix from our data frame, and then, run an EFA by calling these following commands:"*

```
cor <- cor(data, method = "spearman") # First, creating a Spearman correlatio

n matrix.

efa <- fa(cor, nfactors = 1, fm = "minres") # Now, running EFA.

print(efa) # Print the results.
```

3. The efa syntax appears to use the rwa rather than the cor as the dataframe, which also causes confusion.

Corrected, thank you!

Page 14-15

1. It seems the only information used to make the decision regarding unidimensionality is the variance proportion and the scree plot. Previously, made decisions about handling as non-normal data, so was anything further examined regarding assumptions met or not a focus at this stage with this preliminary step before getting into the GRM?
2. Overall, further guidance on making this decision when it is not as clear cut either with some recommended criteria or at least referral to other resources.

Fair point. We have added more nuanced explanations of why it is important to test dimensionality before fitting the model and how to make decisions based on EFA analysis, which we detail later to address your recommendations on later parts below.

Page 16

1. Lines 11-13: this sentence about theta could be clearer here regarding what theta represents again for the readers.

We have mentioned the definition of theta earlier in the introduction, as you suggested in the earlier recommendation (p. 4).

> *"...the covariance matrix of a given dataset to be decomposed into latent constructs (θ) representing true scores or actual skill or trait levels..."*

2. Lines 22-27: seems theta and rwa perhaps are being used to represent more than one thing in how the syntax and descriptions are being written. Theta is called rwa as a construct; the dataset is also rwa. For individuals knew to all this it can be confusing.

Thank you for pointing this out! We have changed this in the manuscript and in the supplementary material. The theta is simply called `theta` and the dataset is now called `data`.

3. Connecting for the readers that the reason assuming the model contains only one theta is because the EFA just did supported that unidimensionality would help connect dots for readers not familiar with these analyses.

Great point. We added it as you suggested (p. 22):

> *"This code implies that we want to estimate a model with one θ namely* `theta`*, which is supported by EFA results we ran earlier."*

Page 17

1. Table 4 … word presentation issue - {apatb-table4} (this happened a couple other times as well)

Thank you for pointing this out! Apparently, there was a formatting error with the previous version of our quarto manuscript, but we have corrected this and make sure that the errors do not appear in the current revision.

2. Explain discrimination for people who are not familiar; as well as how the mean then relates to the interpretation of the b1 values.

We have explained what item discrimination means in the introduction (p. 6):

> *"...fills this gap by allowing item ability to differentiate individuals with varying levels of θ (i.e., item discrimination parameter - a) to differ."*

...and corresponds the definition of item discrimination when interpreting the model's discrimination parameters (p. 22):

"As we see in Table 2, item discrimination parameters of the RWA scale range from high to very high (1.57 - item #1 to 3.32 - item #7), indicating a strong ability of the RWA items to differentiate between individuals with different RWA levels."

3. Last paragraph: connect the points here that you are reviewing for the reader to help them to clear the analysis steps and what is gained or compared across them.
    a. Do both EFA and GRM need to be done?
    b. Can do with same data (this is not the case with EFA and CFA) so helpful to explain why for the reader (increases clarity)
    c. Different variance accounted for by the two approaches. Explain more why this is the case and the interpretation of the two different variances to lean on the GRM to support the unidimensional claim. Stated two different correlation matrices used … explain this more; and why the GRM is better. Could just do with a different matrix input in EFA? (help readers not familiar understand)
    d. Assumed 1 factor, so not actually testing if it is a single factor (as there was not a comparison to a 2-factor model for example). Provide evidence supporting the assumption. If not such a clear-cut case, and there may be more than one dimension could/should the researcher test a 2- versus a 1-factor solution?

Fair point… We agree that these issues are extremely important to cover in the manuscript and we did not have sufficient space to cover it. Since now the editor agrees to grant some discretion on word limit, now we include a more extensive elaboration on what to do if no theoretical argument is present regarding dimensionality (i.e., perform a parallel analysis before proceeding to EFA) on p. 20…

"In a case where there is no or unclear theoretical assumption or empirical evidence of dimensionality, such that researchers are unsure of the θ structure of their item pool, it is suggested that a parallel analysis be performed before proceeding with EFA (O. L. Liu et al., 2007). The purpose of running a parallel analysis is to identify the optimal number of factors underlying the item pool by comparing the eigenvalues of the data with a set of simulated data sets. Parallel analysis can be easily applied by running the following command:

```
fa.parallel(data, fm="minres", fa="pc")
```

As suggested by Guo and Choi (2023), here we use principal component analysis (PCA - `fa="pc"`) to estimate the eigenvalues. When researchers have a Likert scale with five or more but less than 8 response categories, it is recommended to perform a parallel analysis with a polychoric correlation matrix (Guo & Choi, 2023), which can be done by adding `cor = "poly"` to the above command.

After running the code, readers will see in the console that parallel analysis suggests two components underlying the item pool, but this is not a clear cut since the eigenvalue of the second component is only slightly above 1 (i.e., 1.303). In this situation, we can run another EFA to be sure by assuming that there are two factors underlying the data, and then compare the results with the single-factor EFA we ran earlier."

```
efa2 <- fa(cor, nfactors = 2, fm = "minres") # Running EFA assuming that ther

e are two factors.

print(efa2) # Print the results.
```

We also show the result of EFA with 2 factors, and then, give suggestions on how to make a decision based on these results (pp. 20-21).

> *"EFA with two factors has slightly more variance explained and lower root mean square residuals (Cumulative Var = 0.57, RMSR = 0.03) than the EFA we ran earlier (Proportion Var = 0.53, RMSR = 0.05), indicating a better model fit. However, the correlation between the first and second factors is large (r = 0.79). In this case, it is difficult to definitively conclude that our data show a single factor (i.e., unidimensional), but we will proceed with the unidimensional GRM analysis because the RWA theory underlying the item pool argues for unidimensionality, the one-factor EFA shows acceptable evidence of unidimensionality, and, in general, a one-factor model is easier to interpret. It should be noted, however, that we can expect some local dependency problems with our one-dimensional GRM model later, since the two-factor model is a slightly more accurate representation of our data, according to our EFA results."*

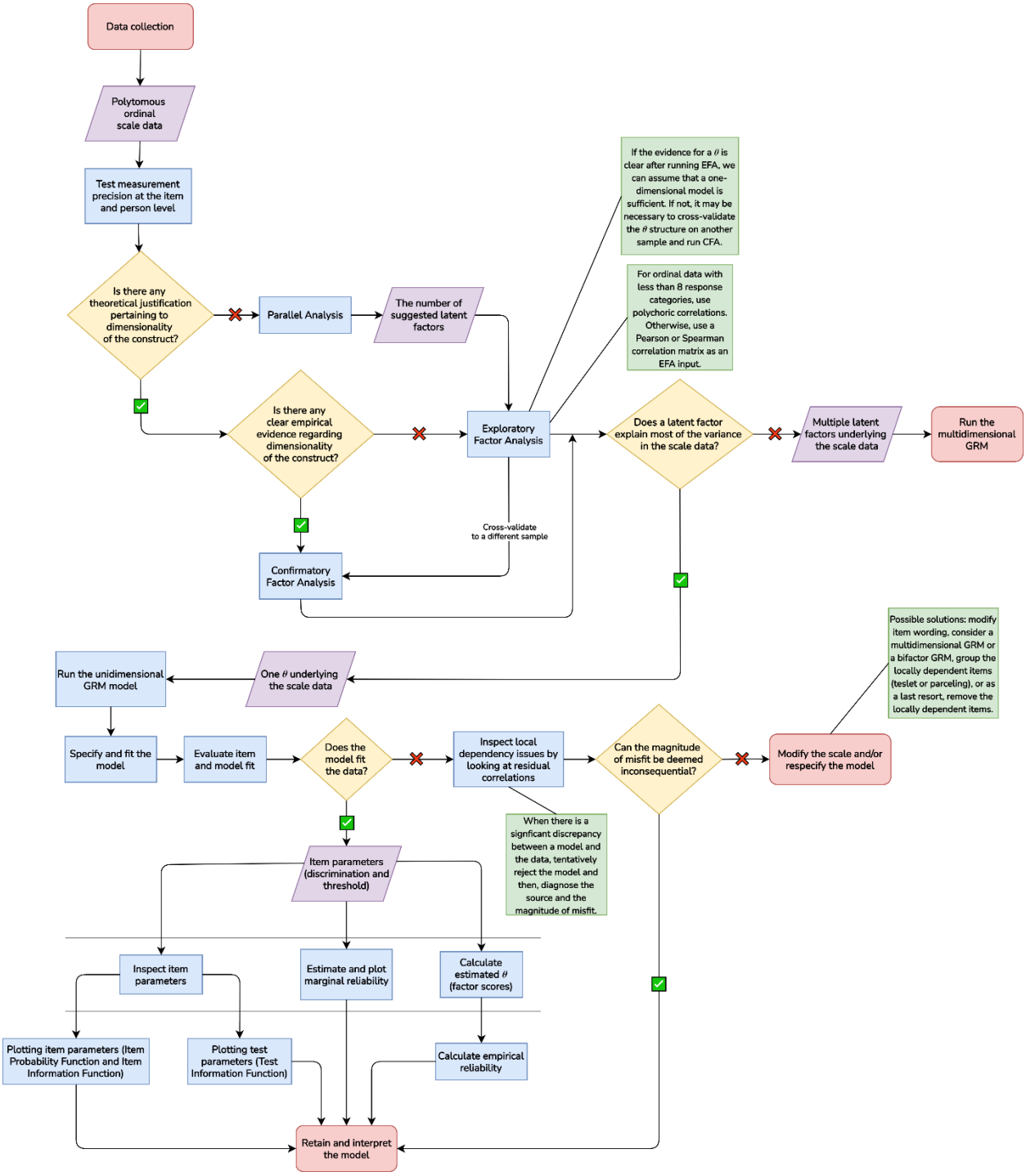
Finally, we also clarify why after running EFA, GRM needs to be done, and in which conditions, performing CFA may be necessary (p. 21):

> *"When stronger evidence of unidimensionality is needed, especially when testing the precision of a newly developed scale, researchers can also cross-validate the θ structure of their scale on another sample by running a CFA (Flora & Flake, 2017), and then decide to run a uni- or multidimensional GRM analysis after obtaining stronger evidence of the number of θ underlying the item pool. Note that parallel analyses, EFA, and CFA are used to uncover the θ structure of the item pool and to focus on the snippet of the performance of the measure at the sample level. If researchers are interested in item-level and person-level parameters, then GRM should still be conducted. EFA and GRM can be performed on the same dataset, since the purpose of each analysis is different, but EFA and CFA should be performed on a different sample, with EFA performed at the early phase and CFA performed at a later stage, after obtaining preliminary evidence of the underlying θ structure (Flora & Flake, 2017)."*

We illustrate the process of testing dimensionality as a part of an annotated flowchart, which we put in the discussion section, as follows:

Figure 7

An Annotated Flowchart Depicting the Process of Fitting A GRM Model



Page 18

Table 2: again, seems to be from the items before reverse coding addressed, because the lambdas are negative for those same items

Thank you! Data cleaning issues have been corrected.

Page 20

1. Explain M2 stats more (way currently written distinguishing M2 stats from goodness-of-fit (again recommend not using acronym when not needed) indices. Relating the M2 role to the chi-square role of CFA may be helpful for readers.

Added as follows (p. 26):

> *"M2 statistic is conceptually similar to the χ2 test of overall model fit in SEM (Maydeu-Olivares & Joe, 2006), so the interpretation is similar."*

2. There are excellent or close, good, and acceptable criteria for goodness-of-fit indices with CFA (Little, 2023), so it is recommended that guidance be presented that is clearly specific for GRM.

We clarify this by explaining that the cutoff criteria are for a good fit and give more information about the cutoff criteria for an acceptable fit. In general, the interpretation of fit indices is the same for CFA and GRM (p. 26):

> *"The rule of thumb for other goodness of fit statistics is also similar to a general guide for interpreting goodness of fit in CFA or SEM models (Schermelleh-Engel et al., 2003), which is RMSEA ≤ 0.05, SRMSR < 0.05, TLI and CFI ≥ 0.97, for a good fit. While 0.05 < RMSEA ≤ 0.08, 0.05 < SRMSR ≤ 0.10, 0.97 < TLI and CFI ≤ 0.95, for an acceptable fit."*

3. Provide more specific p-value for the M2 statistic.

The p-value of M2 statistic is very small, so we change it to (p. 26):

> "M2(209) = 1,980.25, $p < .001$"

4. Be more clear regarding the model fit interpretation. According to above criteria, the RMSEA did not meet good fit; SRMR was almost good fit; and TLI and CFI were essentially .97

Thanks for the correction. We explain more about the model fit interpretation on pp. 26-27 with the hope that it now captures the nuance of interpreting model-data fit, and helpful for the readers.

> *"According to our analysis, the model does not fit the data well (M2(209) = 1,980.25, p < .001, RMSEA = 0.098, SRMSR = 0.055, TLI = 0.969, CFI = 0.972), because there is a significant discrepancy between response patterns predicted by the model with the data, as shown by significant M2 statistics. When faced with this situation, readers are advised to tentatively reject the model and then identify the sources and magnitude of the misfit (Kline, 2023; Y. Liu & Maydeu-Olivares, 2014). It is likely that a model fails the χ 2 test, but the residuals show an*

insignificant discrepancy between the observed and expected values when we have a very large sample (Kline, 2023; Maydeu-Olivares, 2015), which may well be the case since we have more than 9,000 cases in our dataset.

Overall, our model is neither a poorly fit nor a good fit. RMSEA shows a poor fit, while CFI and TLI indicate a good fit. However, the most important goodness-of-fit statistics we should pay attention to is SRMSR, since it provides the average effect size of model misfit (Maydeu-Olivares, 2015). SRMSR is only slightly above the 0.05 cutoff, suggesting that some areas of our model may be misfitting, but the misfit is not so severe. With this in mind, it is noteworthy to identify the source of the misfit, and local dependency between items is the main suspect."

5. Introduce and explain the signed chi-square statistic more (again note how appears in text).
6. Provide interpretation for the RMSEA at the item level

We have rewritten the interpretation of item-fit statistics as follows (p. 27):

"The output of this function is provided in Table 3, which shows the signed χ2 for item fit (Orlando & Thissen, 2000) for each item. All items have very good RMSEA values (≤ 0.05), indicating that each individual item fits the underlying IRT model. Normally, we also want the p values of the signed χ2 to be insignificant (p > .05), suggesting that there is no discrepancy between the observed response pattern and what the model predicts. In Table 3, however, only three items here are not significant (#2, #4, and #12), which is expected when we have a very large sample (Kline, 2023). Nonetheless, it may also be the case that some of these items are not locally independent. We will explore this issue further in the next part."

Page 21-22

1. Table 3 not table 5 … why is the df changes for each item?

Items do indeed have different dfs because each item has different response distributions. The df for the S-X2 test is computed as the number of rows remaining after collapsing (to ensure sufficient expected cell frequencies) minus the number of parameters used to calibrate the item (Orlando & Thissen, 2000). When the data is sparse, which means there are relatively few responses in certain categories, some score groups may need to be collapsed or combined to ensure that the expected frequencies in each group are sufficiently large. This collapsing reduces the number of score groups available for the S-X2 test, which in turn lowers df for that specific item.

2. All the items appear to have a significant p-value (include 02, 04, and 012). I presume these are for the signed chi-square by df, so it would seem better to have the p-value next to those columns rather than the RMSEA. Otherwise, include clarification regarding this.

There was a formatting error in our previous manuscript. Now it is corrected, and indeed there are three items with nonsignificant S-X2 p value. The order of column is corrected as well (p. 28).

Table 3

Item Fit Statistics

Item	S-χ2	df	*p*	RMSEA
001	1,187.23	957	0.0000	0.005
002	917.58	915	0.4698	0.001

Item	S-χ2	df	*p*	RMSEA
003	983.47	847	0.0008	0.004
004	978.46	930	0.1313	0.002
005	1,072.31	957	0.0054	0.004
006	1,310.80	1,041	0.0000	0.005
007	972.91	780	0.0000	0.005
008	1,191.41	1,041	0.0008	0.004
009	1,254.66	981	0.0000	0.005
010	981.24	770	0.0000	0.005
011	1,206.09	988	0.0000	0.005
012	842.01	811	0.2187	0.002
013	1,088.69	904	0.0000	0.005
014	1,031.66	903	0.0018	0.004
015	1,152.22	932	0.0000	0.005
016	997.76	891	0.0071	0.004
017	1,142.47	906	0.0000	0.005
018	1,182.62	936	0.0000	0.005
019	912.71	793	0.0020	0.004
020	1,262.10	1,046	0.0000	0.005
021	1,031.48	878	0.0002	0.004
022	1,122.72	832	0.0000	0.006

Note. Signed χ2 Statistics. RMSEA = Root Mean Square Error of Approximation, CFI = Comparative Fit Index, TLI = Tucker-Lewis Index, SRMR = Standardized Root Mean Square Residual.

3. Given, the sample size and large model by the df having a significant p-value seems rather likely; so some more context for interpretation taking into account such information is helpful.
4. The suspicion of not being locally independent seems odd wording given the items are all measuring the same construct, it is likely that there is additional variance different items share that is not part of the overarching RAW construct.

Great point! We add this in the manuscript (p. 27):

> "*In Table 3, however, only three items here are not significant (#2, #4, and #12), which is expected when we have a very large sample (Kline, 2023). Nonetheless, it may also be the case that some of these items are not locally independent. We will explore this issue further in the next part.*"

Page 23

1. the description of Yen's 03 statistic could be increased to explain why this option over the other, and are there exceptions to this recommendation based on data type, sample size, etc?

We modify the suggestion of choosing the technique detecting residual correlations as follows (p. 30):

> "*To test whether the item pool is locally independent, mirt provides several alternatives for examining the behavior of the residuals. The first option is to run the Local Dependency (LD) χ2*

statistic (Chen & Thissen, 1997), which looks at the covariance between pairs of items after accounting for θ. In this tutorial, we demonstrate the use of Yen's Q3 statistic (Yen, 1984), which is deemed more powerful for detecting underlying local dependency in unidimensional models, especially when dealing with polytomous data, than the LD χ2 statistic (Chen & Thissen, 1997). For comparison, we also demonstrate the use of LD χ2 statistics in the annotated R Markdown file."

2. explains why .2 is the absolute minimum cutoff used and provide reference support for this (again helping readers who may run this analysis in the future). You can see this with residual correlation information with other analyses, such as CFA, so highlight the importance of this information and Line 48: sentence starts with interestingly. It seems these results should align, so not sure why interestingly is used. More interesting to me is that there is not a clear pattern regarding why these. For instance, they are not all the reverse coded items, so that isn't what is being picked up on for the residual correlation.

We cannot really add more than mentioning the rule of thumb, given the tight word limit, but we're including the reference now (p. 30) so that readers who are curious about why |0.2| is the critical value of the Yen's Q3 can read further in the cited paper.

"As a practical guideline, Chen and Thissen (1997) suggests that a critical value for the Q3 statistic is around |0.2|, so we are interested in item pairwise correlations above |0.2|."

3. Providing further insight into why these items are related to each other is helpful.
4. Also, connecting to when the next step examination of this and whether these items have shared characteristics and what do about that is important.

While it is potentially interesting for readers to look closely at the items to see why they are significantly correlated even after accounting for theta, we refrain from adding more text at this point because the current manuscript exceeds the word limit by more than 2,000 words, and we are committed to using the editor's discretion judiciously. We feel that this type of discussion would be more suitable for a paper focusing on the RWA construct, while our paper here uses the RWA scale only as a case study. We do, however, recommend that readers look first at the strong locally dependent item pairs and then at the remaining locally dependent pairs (pp. 30-31). We have included the strategies for dealing with local dependency in our discussion section.

"After running the code, readers may see in their R console that the residual correlations between items #3, #4, #5, #7, #11, #13, #14, #18, #19, and #21 are above |0.2|. As suspected, all of these items, except #4, also have significant S-χ2 statistics (see Table 3), which tells us that we need to look at these items closely and then decide whether these items have common characteristics beyond what is explained by the model. While it is worth looking at all of these item pairs, we recommend that readers prioritize the investigation of severe local dependencies, as indicated by item pair correlations above |0.3| (i.e., #14 and #3, #11 and #20, #5 and #1, #18 and #2), before addressing the rest of problematic pairs."

The title can have spaces in it? R usually does not like spaces

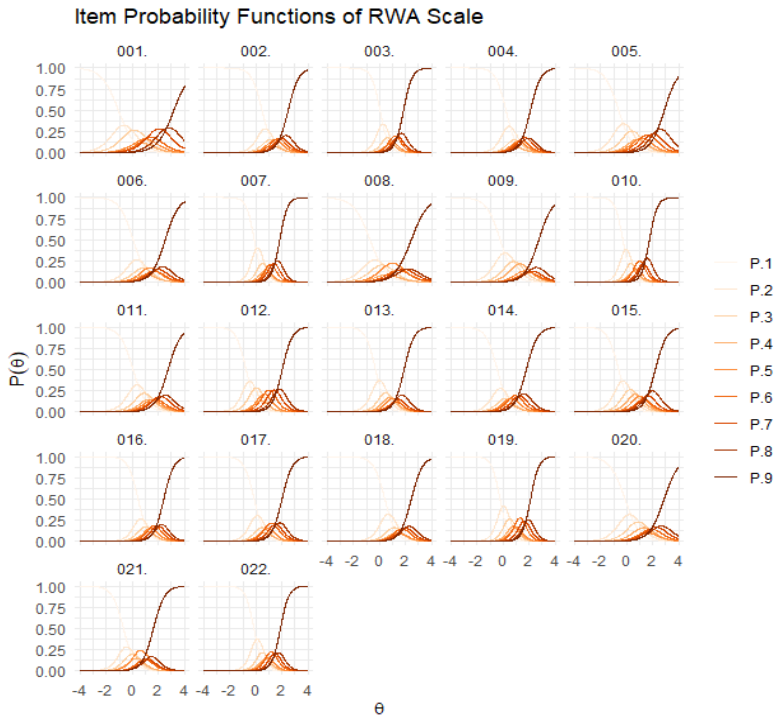
Provide more information about interpretation of these figures. For example, walk through a couple with different patterns to help the reader understand how to actually interpret these and the information they can glean from them. How is what they saw for example in table 2 being illustrated in figure 3?

We now add more lines to explain what the figure means, as follows (pp. 31-34):

> *"The output of this function is Figure 3, which we can see that as θ increases, participants are more likely to choose higher response categories. The gradual transition from one response option to the next across the spectrum of θ values indicates that the items are capturing incremental increases in the RWA level, which is consistent with the gradual increase in threshold parameters we saw in the previous step. However, all IPFs of RWA items seem to be significantly overlapping and tend to peak on a θ value close to or higher than the mean. This implies that RWA items are more sensitive to differentiate participants with high levels of RWA.*

Figure 3

Item Probability Functions of RWA Scale



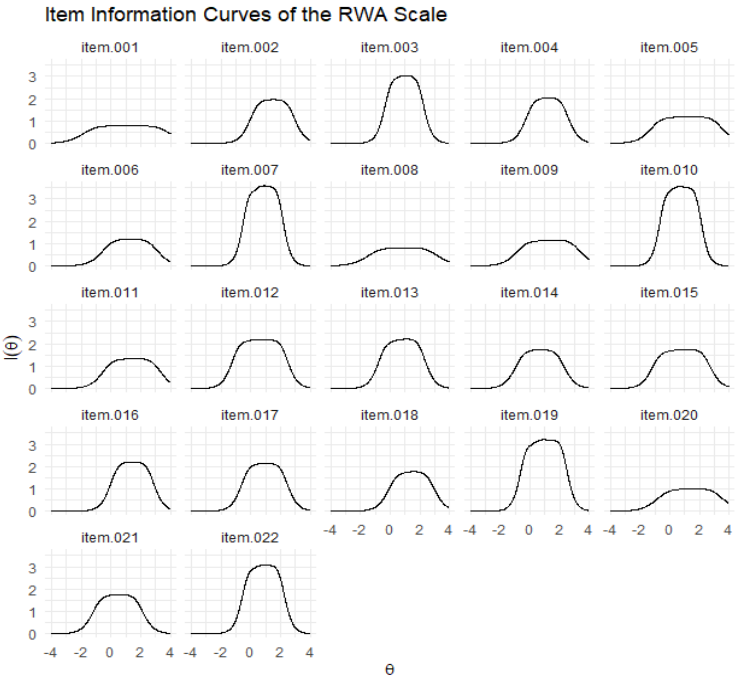
To evaluate the performance of each item in measuring the RWA trait, we can visualize the amount of information explained by each item with this simple line of code:

```
itemInfoPlot(fit, facet = TRUE, title = "Item Information Functions of the RWA Scale")
```

The output for this function is Figure 4 where the x-axis represents the range of θ, and the y-axis indicates the amount of information provided by each item. Peaks that are higher and narrower indicate items that are very informative for specific levels of θ. In our case, most items provide the highest amount of information near the center to the right side of the θ distribution, indicating they are most useful for individuals with average to high levels of RWA."

Figure 4

Item Information Curves of the RWA Scale



...and we hope that they are sufficiently clear for the readers.

Page 25

1. Again, it looks like the items were used pre-reverse coding because the probability functions are flipped for the negatively worded items.
2. Also, if you change the item names to 001-009, 010 ... then they will appear in numerical order, which will help people trying to connect the figures with the information in the tables.

Thank you very much for the tip! As you can see above the scoring has been corrected, and the IPFs and IIFs are ordered by item number.

Page 26-27

1. Suggest not using acronyms when talking about the different functions. It was good to introduce for people to know they can be referred to that way; however, to help people knew to this material, I would recommend writing them out to remember the differences. Use one or two items to illustrate how to interpret. For example, item 1 or 11 and item 10 have very different distributions.

We appreciate the feedback, but the acronym for these figures appears only a few times in the manuscript, and only in the same section ("Step 5: IRT Plot"), so keeping them short is our decision at this time, especially when we are desperately trying to keep the manuscript short and concise.

2. Also, the figures have x-axis labels for items 5-9 only

Corrected, thank you!

Page 28

1. Explain the information gleaned and interpretation more. The x-axis is going from -2SD to +4SD. That's uneven so what does that tell us? It is accounting for 27.5% below the mean and going beyond 49.85% above the mean. There are very few people outside of this range.
2. There are two values on the y-axis, so letting people know what this represents is helpful. Also, did not reference the SE information that is in the graph in the manuscript; however, this is relates to the point regarding at least the below the mean not as good for measuring those beyond -2SD; certainly -3SD

We add a few words mentioning the SE (p. 37) but we feel that the explanation is already clear, so we don't add more explanation beyond what is already there in the manuscript.

> "*The output of this code is Figure 5, which we can see that the RWA scale as a whole is informative for measuring a group of individuals with a wide range of RWA levels, i.e., between -2SD (low) to +4SD (very high). Moreover, the RWA scale is not optimal for measuring individuals with RWA levels below or beyond this range, as we can see in Figure 5 that standard error increases below -3SD.*"

Page 30-32

1. Define each of the reliabilities (marginal reliability, reliability #2, and empirical reliability) for the readers, why it is important, and how to interpret it (criterion or guidelines)

The information regarding theta distribution of each reliability has been there already, for example on page 38:

> "*The output of the above code is 0.948, which indicates that the overall RWA scale is reliable, assuming that the underlying θ distribution follows the Gaussian or normal distribution.*"

...and page 40:

"mirt can also compute the overall reliability of the RWA scale, assuming the model-predicted θ distribution (i.e., factor scores) and its associated standard errors."

2. Explain further regarding different distributions and which yours falls into to help with interpretation and readers future implementation.

We think this is pretty clear already as we have argued in your previous recommendations.

3. What is the different reliability that is being examined with the second code and the difference in it versus the marginal reliability? What is the information gained for each and importance of examining both?

We believe that this has been very clear in the manuscript. We have mentioned that marginal reliability assumes the normal θ distribution, and can be different across different levels of θ, as follows (p. 37):

"The output of the above code is 0.948, which indicates that the overall RWA scale is highly reliable, assuming that the underlying θ distribution follows the Gaussian or normal distribution. Since marginal reliability can vary across different levels of θ, readers can also visualize the marginal reliability given the levels of θ by running the function of ggmirt package, as follows..."

...while empirical reliability assumes the model-predicted θ distribution, and applies as a general reliability across all participants in the sample (p. 39).

"The output of these codes is 0.953, which is rather close to marginal reliability we estimated before. This implies that RWA scale is overall highly reliable in measuring RWA trait across all participants in the sample."

4. Is .75 the recommended criterion?

We correct this, as follows (p. 37):

"...which shows that the RWA scale can measure individuals with θ levels between approximately -1.75SD and +2.75SD with optimal reliability, i.e., rxx ≥ 0.80."

5. Connect this information with the prior results both reliability and the SD range in the prior section

Good point! We add this already at the very end of the "Step 6" (p. 39):

"Generally, reliability analysis aligns with the IPFs and IIFs we saw in the previous step, i.e., the RWA scale is highly reliable overall, but works best to measure the RWA construct across participants with RWA levels between ~2SD below the mean and ~3SD above the mean."

6. Also, in figure 6, -2SD is not at .75, so increased clarity regarding interpretation and criterion is necessary.

Corrected, thanks! (p. 37)

"...the RWA scale can measure individuals with θ levels between approximately -1.75SD and +2.75SD with optimal reliability, i.e., rxx ≥ 0.80. This means that the RWA scale is most reliable

for participants with low to high levels of RWA, but it becomes less reliable when measuring participants with extremely low (θ < -1.75SD) or extremely high levels of RWA (θ > 2.75SD)."

Page 32

Explain more what is captured with this estimated theta and the SE. These are important but presented very briefly. Recommend again including the values and interpretation of one or two items.

Thanks for suggesting this! We add a few lines what is estimated theta, as follows (p. 39):

"*If item discrimination (a) and thresholds (b) we showed earlier reflect item parameters, then the estimated θ or factor scores we compute here are the person parameters, since they represent each participant's "position" on the θ continuum. To compute empirical reliability, one needs to calculate the estimated θ value, and its associated standard error, of each participant first, and then, calculate empirical reliability based on these values.*"

Discussion, page 33

1. "briefly explain how IRT differs from CTT in its assumptions" did not really present this which should have been part of the introduction as noted above.

We add a new section in the introduction ("A Brief Overview of Item Response Theory), as we have explained at the beginning of your feedback.

2. The fit of the model seems acceptable not poor (line 28) based on what was presented and described earlier.

Fair point… We rewrite the sentence to (p. 42):

"*Note that the model we estimate does not fit the data perfectly due to local dependency, as we showed in steps #3 and #4.*"

3. More discussion on the handling of local dependence, such as parceling (Little, 2013), correlated residuals included in model, subscale constructs, and bifactor (general and specific constructs). Some of these address and represent the information differently, so further discussion on the issue of remaining correlations and why different potential solutions.

We add more potential solutions to model misfit, including parceling locally dependent items as you suggested (p. 42):

"*If researchers encounter this problem when evaluating their scale data, we recommend tentatively rejecting the model and then locating the source of the model misfit before interpreting measurement precision or drawing substantive conclusions based on these estimates (Kline, 2023). In doing so, researchers can take a closer look at residual correlations to identify the source and magnitude of misfit (Kline, 2023; Maydeu-Olivares, 2015).*

Once the sources of misfit are identified, one solution to address them is to rephrase or combine the content of the problematic, locally dependent items and then cross-validate the modified scale on another sample. Alternatively, researchers may need to reconsider the θ structure. If the data are unidimensional, but some of the items are locally dependent, this may indicate the existence of multiple θ within the data structure. In this sense, modeling the test data as a correlated multidimensional model or a bifactor model (i.e., multidimensional models with a g factor) may be a viable solution. Another solution to consider is to group locally dependent items by modeling them together as a "testlet" (Cook et al., 1999), or to combine locally dependent items into a single composite score (i.e., item parceling), especially when the goal of the analysis is to understand the construct being measured rather than to identify the relationship between items and the θ (Little et al., 2013). Removing locally dependent items should be a last resort, as it may improve model fit, but can jeopardize measurement precision."

Outlook

FW: IJP-REA-23-453.R2 Decision Letter

From Zein, Rizqy Amelia <Amelia.Zein@psy.lmu.de>
Date Tue 11-Mar-25 07:42
To amelia.zein@psikologi.unair.ac.id <amelia.zein@psikologi.unair.ac.id>

-----Original Message-----
From: David Giofré <onbehalfof@manuscriptcentral.com>
Sent: Friday, 18 October 2024 12:07
To: Zein, Rizqy Amelia <Amelia.Zein@psy.lmu.de>
Cc: abigail.gewirtz@asu.edu; david.giofre@unige.it
Subject: IJP-REA-23-453.R2 Decision Letter

18-Oct-2024

RE:

MS number: IJP-REA-23-453.R2
Title: Getting Started with the Graded Response Model (GRM): An introduction and tutorial in R
Authors: Zein, Rizqy; Akhtar, Hanif

Dear Mrs. Zein,

Thank you for taking the time to revise and resubmit your manuscript. I read through your paper and letter in response to the reviews and sent it out to the previous referees for their comments. The review comments are included at the bottom of this letter.

As the reviews suggest your paper is now suitable for publication, I am happy and pleased to be able to accept your paper in its current form. Once our managing editor has ensured it meets all our pre-production requirements, your manuscript will be forwarded to the publisher for copy editing and typesetting. You will later receive proofs for checking and instructions for transfer of copyright. Please ensure that all your figures are of high resolution, as this will enhance the readability and impact of your report.

Within a few days of receiving your manuscript, Wiley's Author Services will send the corresponding

author an email asking them to log in to the IJP system where they will be presented with the appropriate license agreement for completion. Please note: your article cannot be published until a signed license agreement has been received.

Thank you for your contribution to the International Journal of Psychology. We look forward to receiving further submissions from you in the future.

Sincerely,
David Giofrè
International Journal of Psychology

P.S. – You can help your research get the attention it deserves! Wiley Editing Services offers professional video abstract and infographic creation to help you promote your research at www.wileyauthors.com/eeo/promotion. And, check out Wiley's free Promotion Guide for best-practice recommendations for promoting your work at www.wileyauthors.com/eeo/guide.

Reviewer(s)' Comments to Author:

Reviewer: 1

Comments to the Author
Thank you for the opportunity to review a revision of the manuscript. I felt strongly this would be a good addition to the literature if the authors made a few key changes.
I believe the authors did an excellent job responding to my comments and highlighting the utility of their tutorial. I feel satisfied with the changes they did and did not make based on my feedback. I do not have further concerns to express, however, I do want to call out Fig. 7 specifically. This figure is fantastic and has significant utility for people interested in GRMs. I could see it being referenced in course materials on model selection.

Reviewer: 2

Comments to the Author
Thank you to the authors for their clear investment in this manuscript and for addressing the comments from the reviewers. I am glad you were allotted the additional space needed for the added specifics, including the brief introduction for comparing GRM to CTT and IRT, along with the accompanying table, which I believe will greatly help readers and increase the overall impact of this manuscript. The added flowchart is also a very valuable contribution and resource for individuals looking to start implementing this analysis. I appreciate the effort and work you put into the manuscript to address the prior recommendations thoroughly.

All papers published in International Journal of Psychology are eligible for Panel A: Psychology, Psychiatry and Neuroscience in the Research Excellence Framework (REF).