Dear reviewer,

Thank you so much for your positive comments and invaluable feedback. The text below in bold is our response to your concerns and questions. In addition, we have highlighted all changes to the manuscript with blue color.

1. Could you give an example of which scenario one would be interested in comparing different models, and how the data will be generated in this case?

**Thank you. We have added the following paragraph in section 3 discussing simulation studies in IRT.**

**“Simulation studies have been widely used to compare IRT models in various ways. Some simulation research has used one model to simulate data and fit the simulated data to one or more candidate models (Lee & Bolt, 2018a, 2018b; Molenaar, 2015). These designs are appropriate to test the accuracy of software or estimation methods (one fitted model) or to investigate the extent to which multiple fitted models can recover features of the data-generating model. Another simulation design generates data according to two or more models and fits those data to one or more models (e.g., Fujimoto and Falk, 2024; Kang and Cohen, 2007; Zhang et al., 2022). This design is typically used to compare the relative accuracy or efficiency of different models, and is often used in model selection research. The methods proposed in this study are relevant to simulation studies that follow this second design.”**

1. As mentioned, the ICOMP model selection criterion can also be used when complexity needs to be considered.

**Thank you for this suggestion. However, to the best of our knowledge, the ICOMP criterion has not been tested for IRT model comparison and it is not available in the most commonly used software for IRT simulation research. In section 1, we added the following text:**

**“However, to the best of our knowledge, the MDL and ICOMP methods have not been tested for IRT models and are not available in popular IRT software packages such as mirt (Chalmers, 2012). As such, the analyses in this paper use two widely-used model comparison indices, AIC and BIC.”**

1. It would be helpful if the authors clarified the purpose of their proposed methodology (e.g., it is proposed for simulation studies with different models….) and gave examples when this will be the case.

Section 3.1: What you aim to do here in conditions a) and b) is to generate parameters from equivalent distributions. The reason for doing that needs to be better explained in the introduction.

**We address these two pieces of feedback together, as the purpose of the paper is closely related to what we did in section 3.1. We discuss this at the end of section 2 and the beginning of section 3:**

**“The 2PL model and the LPE model have similar link functions but different parameter counts, and the 2PL model and the NLL model have the same parameter counts but different link functions. Therefore, to investigate the influence of different function forms, we will compare the 2PL model and the NLL model in Study 1; to investigate the influence of different parameter counts, we will compare the 2PL model and the LPE model in Study 2.”**

**“Simulation studies have been widely used to compare IRT models in various ways. Some simulation research has used one model to simulate data and fit the simulated data to one or more candidate models (Lee & Bolt, 2018a, 2018b; Molenaar, 2015). These designs are appropriate to test the accuracy of software or estimation methods (one fitted model) or to investigate the extent to which multiple fitted models can recover features of the data-generating model. Another simulation design generates data according to two or more models and fits those data to one or more models (e.g., Fujimoto and Falk, 2024; Kang and Cohen, 2007; Zhang et al., 2022). This design is typically used to compare the relative accuracy or efficiency of different models, and is often used in model selection research. The methods proposed in this study are relevant to simulation studies that follow this second design.”**

1. Another issue that I would like the authors to discuss is that changing the link function changes the scale, and therefore, the parameters will also be on a different scale.

**Thank you for raising this issue. We have added the following paragraph to Section 3:**

**“It is important to note that this method of comparing different IRT models assumes that all fitted models are on the same latent trait scale. However, it is often unlikely that this will strictly be true. For example, Shim et al., 2024 noted that the one-parameter NLL is a nonlinear transformation of the one-parameter logistic model. In other words, these two models are theoretically capable of making the exact same predictions as each other, corresponding to different underlying latent**

**trait distributions. However, in practice, these models are often identified by specifying a standard normal prior for θ during model fitting, hindering the ability of the different models to be exact transformations of each other. Moreover, much IRT model comparison research, including that presented in Shim et al., 2024, does not explicitly take this potential confounding factor into account, and it is unclear how large the effects of scale differences may be. As such, we acknowledge that**

**differences in scale are a potential confounding factor that is not accounted for in the following studies, but we still find value in controlling for the IRF features that we do control for.”**

1. In the discussion of Figure 1, the left panel should be right, and the right panel should be left. Furthermore, different maximum slopes are expected because the models are different, and even the interpretation of differs in each model based on the link function used.

**Thank you for correcting! We have changed the “left” and “right” in the text.**

**We also agree with your second sentence. This is an illustration why we want to control IRFs instead of parameters when generating data from different models, because the same parameters may lead to different max slopes and corresponding locations.**

1. In section 3, condition c) adjusted NNL that mimics the properties of a). What do you mean by mimicking the properties of a)? In what way does the NNL mimic the 2PL?

**By mimicking properties of a), we meant that the highest slopes and the corresponding s of the IRFs generated by the NLL model mimic those of the IRFs generated by the 2PL model. We wanted to control their IRFs as closely as possible.**

**We have changed the phrase’ mixmics the properties of the…” to “matches the slope and location properties of the…” and added teh following sentence to Section 3:**

**“ In this way, the IRFs generated by the adjusted 2PL model should be more similar to the IRFs generated by the NLL model than those from the unadjusted 2PL model.”**

1. In study 2, the LPE is known to be challenging to fit in small samples or generally?

**Both, because compared to the 2PL model, the LPE model needs a larger sample size to have a stable estimation, and in general, the high intra-correlation makes it hard to estimate.**

**“a consequence of the known difficulties associated with fitting this model, resulting from the high within-item correlation among its parameters”**

1. ‘We compared the fitted models…’ The RIMSE is computed for each model separately and then compared across models. Maybe you need to explain that.

**We have added the following text to Section 3:**

**“We calculated the RIMSE separately for each fitted model and compared the**

**distribution of RIMSE values across different data-generating and fitted model**

**conditions.”**

1. Section 3.1: Is it necessary to check the parameter recovery of 2PL and NLL models when data are generated from the two models, respectively? Don’t we already know that?

**Yes, we may expect that IRF recovery should be better for the data-generating model. However, including this information is an important aspect of our research. This section shows that the extent to which a misfit model has worse recovery than the correct model depends on how exactly the data are generated. We would not be able to demonstrate this without including the parameter recovery analyses that we do.**

1. Section 3: What is a penalized goodness-of-fit?

**By this term, we refer to the AIC and BIC information criteria.**

**“To evaluate penalized goodness-of-fit, we used the AIC and BIC because they penalize goodness-of-fit by parameter counts.”**

1. Can you add in the text that in the left panel of Figure 2, data were generated by the 2PL, NLL, and aNLL models, but the model fitted to all those data was the 2PL model? If this is the case, then I am not sure I can follow the following: ‘However, these plots illustrate that different sets of data-generating parameters can yield different sets of results for the same data-generating model’.

**Yes, we have clarified in the plot notes to Figures 2 and 3 that different colors refer to different data-generating models.**

1. Below Table 2, it is stated: ‘In all cases, the data-generating model was selected significantly more often when the data were generated according to models with higher slopes’ Can you provide a further explanation of that? What does a higher slope imply that makes the correct model be chosen?

**We have added the following sentence to help place this finding in context.**

**“This finding coincides with previous research (Lopez Rivas et al., 2009) demonstrating that high-discrimination items tend to make better anchor items for differential item functioning analysis, possibly because these items provide more information and lead to better model identification.”**

1. Table 2: Can you clarify which models are compared and which data-generating models are in each column?

**Yes. We added another note to Table 2.**

**“2PL", "a2PL", "NLL", and "aNLL" indicate the data-generating model.”**

1. The paper focused on the logit, negative log-log, and logistic positive exponent models. How often have the other models been used besides for the logit model?

**As far as we know, the NLL model and the LPE model have not yet been implemented in operational assessments, however, these models belong to the burgeoning research topic of asymmetric IRT (e.g., the BJMSP journal is assembling a special issue on this topic). We believe that our findings and methodology are increasingly relevant as non-logit functional forms become more widely known and used.**