Reviewer I:

1.) P. 2 Lines 32-35. It is unclear why the multilevel model is explicitly modeled using the WF approach.
Thank you for this remark. Unfortunately, we need too much space to fully explain this in the introduction. We hope that the reader will be patient and we now refer to paragraph 1.3 on P. 2: *"as will be explained in Section 1.3 and 1.4".*

2.) In the introduction, I would recommend spelling out what makes the WF approach for discrete data different from the WF approach for continuous data (Bauer, 2003; Curran, 2003; Mehta & Neale, 2005). Some texts on p. 11 can be under the introduction section to motive the present study earlier.
Thank you for your suggestion. We now added sentences explaining the potential advantages of the WF approach for discrete data in the introduction (see second last paragraph on P.2):
"As will be explained in Section 1.4 and the illustrative dataset, the WF approach turns out to be much more computationally efficient compared to the often used MML estimation method in the LF approach. Using the WF approach does not require specialized multilevel software, which decreases the complexity of estimating multilevel models."

3.) P. 4-5; p. 6. "N" is to indicate a sample size for asymptotic property on p. 4 and "I" was used to indicate a total number of individuals on pp. 4-5. Please use the consistent notation if "N" is "I". On p. 6, "n_j" is used to indicate the number of individuals for a cluster "j". If the authors want to use "I", it should be "I_j". To summarize, the authors need to use a consistent notation.
Thank you for noticing. We chose to use "I" for the total sample size and adjusted all the other formulas accordingly (e.g., Equation 20-22 on P.7).

4.) P.5: Eq. 14. In the third term ("y_i"), please add a summation from $i = 1$ to I or drop an "I" in "y_i". In addition, in "$f(\eta;\theta)$", "θ" is a set of parameters for the prior distribution "η". Then, a different notation for the parameters from all parameters for the likelihood ("$f(y_i|\eta;\theta)$") may be needed.
Yes, we agree with you. We now use "$g(\bullet)$" is the prior density and "$f(\bullet)$" for the likelihood. In addition, we now use $\boldsymbol{\theta}_\eta$ to denote the parameters of the prior only, and $\boldsymbol{\theta}_y$ for the remaining model parameters. To improve the readability of the formulas, we adjusted the formulas of both the PML estimation method (see P. 4-5) and the formulas for the MML (see P.6 and P.8).

5.) P. 8: Eq. 24. "η_y" was not defined in texts. In Eq. 24, one "eta" should be a within-level latent variable and another "η" should be a between-level latent variable. In Equation 24, it may be more precise to present nested integrals (see Rabe-Hesketh et al., 2004). By the way, in Figure 1, the authors use "eta_{fw}" and "eta_y". On p. 14, there were referred as "eta_{fw}" and "eta_{fb}".
Thank you for these comments.
- We now emphasized the explanation of "η_y" better on P. 8 L.3: *" In the multilevel context, the random intercept can be written as a vector of latent variables, referred to as $\boldsymbol{\eta}_y$, so that the extended set of latent variables equals $\boldsymbol{\eta}^* = (\boldsymbol{\eta},\boldsymbol{\eta}_y)$. This formula can be extended to models with more latent variables at different levels of the multilevel model. For a two level model, the log-likelihood for the data can be written as the sum of the log-likelihoods of all the clusters j "*
- Instead of using nested integrals with the MML estimation method, we prefer the notation of Hedeker & Gibbons (1994) and Molenbergs & Verbeke (2005) with a multivariate integral.
- Figure 1 now becomes Figure 2 and uses "eta_{fw}", "eta_{fw}", and "eta_y". We now further explain this on second paragraph of P. 10: *Figure 2 shows a one factor model according to the parameterization of Metha & Neale (2005) for the case where we have three units in each cluster. The within factor is referred to as η_{fw}, the unit-specific version of the variables on the between-level is referred to as η_y, and the between factor is referred to as η_{fb}.*

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. Biometrics, 933-944.
Molenberghs, G. & Verbeke, G. (2005). Models for Discrete Longitudinal Data. New York:

Springer-Verlag.

6.) The following requires more detailed illustrations: "A random-intercept multilevel model can then be estimated as a restricted CFA model, where univariate multilevel models become multivariate uni-level models.
Thank you for this important remark. We added Figure 1 to make this particular sentence more clear. Figure 1 shows the data, the model, and the Formula of a random intercept model in both the LF approach and the WF approach. On P. 9 we refer to this picture: *" Figure 1 shows data, a graphical representation, and a formula of a random intercept model in both the LF approach and the WF approach."*

7.) Pp. 12-13. P. 14. ICC calculation for discrete data is different from that of continuous data in that variance of "y*" is included in the total variance calculation (e.g., see Hox et al, 2017 for binary data; $\pi^2/3$ for a logit link or 1 for a probit link). However, the authors did not describe the calculation of ICC for discrete data. On p. 14: Line 23, I don't follow the calculation of ICCs. Also, it is not common to have the same ICCs for all items in real data analyses.
Thank you for bringing this to our attention. We pulled apart the calculation of the ICC to explain each part of the calculation in a footnote on P.14. The theta parameterization with a probit link is used in our simuliation study. The ICC's we chose is considered 'common' in, for example, educational data (Snijders, 1999). We used different conditions and chose to keep the ICC similar in all items. We felt that different ICCs for different items would make the interpretation of the results of our simulations less clear. This is similar to some other simulation studies (see Jak, Oort & Conor, 2014). However, we are aware that the ICCs fluctuate in real data.

Bosker. A. B. & Bosker, R. J. (1999). Multilevel analysis: an introduction to basic and advanced multilevel modeling. Sage.
Jak, Oort & Dolan (2014) Using Two-Level Factor Analysis to Test for Cluster Bias in Ordinal Data, Multivariate Behavioral Research, 49:6, 544-553

8.) Pp 11-13. A subsection 1.4 includes a proposed approach. I would recommend illustrating each step of the four steps one by one.
Thank you for this very useful suggestion. To illustrate the five step procedure, we added the steps in the syntax of the appendix and refer to this in the text (e.g., P. 10: *"To clarify the proposed steps, we give the lavaan syntax corresponding to Figure 2 in Appendix A."* The R scripts with artificial data can also be found at the Open Science Framework.

9.) P. 13. I would recommend adding explanations why selected simulation conditions can be realistic conditions. Please clarify how many replications the authors considered on p. 13 because the authors explained the calculation SD for SE evaluation.
Thank you for these remarks.
-We now extended the first two paragraphs of the data generation section (Section 2.1) on P. 14, explaining and substantiating the chosen conditions in more detail. However, we do realize that simulation studies cannot represent reality.
-We mentioned the number of replications (500) on P. 14: *"In each condition, 500 datasets are generated. "*

10.) Pp. 15-16. The authors presented parameter recovery of factor loadings only. I would recommend presenting parameter recovery results for all parameters in the model.
Thank you for making this comment. We now plotted the bias of the thresholds in the supplementary materials and we refer in the text to these plots on the first Paragraph of P. 17: *"In describing the parameter estimates, we did not mention the thresholds as these parameters are often not of theoretical interest. For all conditions, the bias in the thresholds was very low for all estimation methods. The figures are shown in the supplementary materials."* All results can also be found at the Open Science Framework. In this way, everyone can study the results.

11.) After reading a section of a simulation study, it is still not clear why the LF approach does not work well for a small cluster size. I would suggest spelling out possible reasons why the LF approach does not work well compared the WF approach in detail based on estimation theory.
Thank you for this remark. We are somewhat unsure what you mean. Despite the (very) small cluster sizes used in our simulation, the MML-LF approach still works very well in general. The exception being the setting where we have a misspecified model. It is unclear to us why MML-LF shows somewhat more bias in this setting, compared to the other approaches (including DWLS-LF). Our study was not designed to critically examine the LF approach in general. Rather, our aim was to show that the WF approach is a viable alternative, in particular when the data is discrete.

12.) I would recommend adding computation time for the LF and WF approaches in the simulation and empirical studies.
Thank you for bringing this to our attention. To give a general idea, we now mentioned the computation time in the illustration on P. 19: *"To give a general idea of the estimation time of the used estimation methods, we saved the estimation time of a model without covariates and errors on the between level; DWLS in the WF approach: 0.11 minutes, PML in the WF approach: 3.24 minutes, MML in the long format approach: 4:25 minutes, and DWLS in the LF approach: 1:53 minutes."* Note that this is a difficult issue as the computation time is dependent on the used software and the implementation of the estimation method. The MML estimation method is the slowest, but the computation time is very dependent on the used number of quadrature points. The PML estimation method can be accelerated in many ways, but this has not been implemented yet. As a result, a direct comparison is difficult.

13.) PP. 19-20. The LF approach can test the restrictions as the WF can, although it depends on software. Do the authors mean that we cannot test the restriction with the LF approach?
No, not necessarily; in addition to restrictions to the specified multilevel model of Figure 6, the WF approach can also test equality constraints across clusters. In Figure 2 the equality constraints (labeled e to h) across clusters then can be tested in the WF approach. For example, units that behave different from all other units in a cluster are easier to identify in the WF approach. We now explain this halfway P. 10: *"The figure shows that the WF approach is more flexible than the LF approach, as the equality restrictions across units in a cluster (e.g., e to h for the factor loadings and/or i to l for the residual variances in Figure 2) can all be tested by freeing the restrictions in the WF approach. Note that within the LF approach, it is not possible to free restrictions across clusters."*

14.) This study focuses on the random intercept model. I would recommend discussing how the WF approach presented in the present study is applicable to the random intercept and slope model.
Thank you for this very useful suggestion. We now added a paragraph in the discussion to reflect on models with a random intercept and a random slope (see last paragraph on P. 21 and the first paragraph on P. 22: *"In this study, we only considered models with a random intercept. Multilevel models can be extended by adding random slopes, where the impact of covariates is allowed to vary across clusters. The estimation of random slopes requires case-wise estimation. With discrete data in the LF, only the MML estimation method can estimate models with a random slope. In the WF approach, the PML estimation method seems best suited to estimate models including a random slope in for example generalized linear mixed models (see Bellio and Varin, 2005; Tibaldi et al., 2007; Cho and Rabe- Hesketh, 2011). Compared to the MML in the LF approach, the PML estimation method in the WF approach can in theory estimate many random slopes and other latent variables. The WLS estimation method in the WF approach uses a two-step estimation procedure and can therefore not estimate models with random slopes."*

Reviewer II:

1.) Adding a discussion of model fit indices in the simulation study would help readers to understand the relative performance of analytic approaches. This is particularly relevant to the correctly- and mis-specified models. How the wide format SEM approach functions under different forms of model misspecification is reasonably outside the scope of the present paper. That said, the authors' simulation study provides some information relevant to this question.
Thank you for this comment. This is a interesting suggestion. In describing the WF approach we explained how to obtain fit measures for both continuous- and discrete data. In the discussion (P. 21 second last paragraph) we discuss the possible difficulties with fit indices with multilevel data. In this study we only focused on the accuracy and efficiency of the estimated parameters. Further research on this topic is needed.

2.) Please comment on recovery of threshold estimates across analytic methods in the simulation study. In addition to accurate estimates of item thresholds, some readers may be interested in the distribution of threshold estimates over 500 trials in each analytic approach.
Thank you for bringing this to our attention. The plots with the thresholds are now added to the supplementary materials and we refer in the text to these plots on P. 17: *"In describing the parameter estimates, we did not mention the thresholds as these parameters are often not of theoretical interest. For all conditions, the bias in the thresholds was very low for all estimation methods. The figures are shown in the supplementary materials."* The results of the simulation can also be found at the Open Science Framework.

3.) The summary of results provides a useful illustration of the mean % bias. It would be very helpful to provide readers with additional information about the distribution of parameter estimates over 500 trials (e.g., through error bars or histograms)
The distribution for the parameters are bell-shaped and approximately normal. Including all this plots in the supplementary materials seems somewhat overkill. This is why the results of the simulation study can be found at the Open Science Framework. The readers can produce the error bars and/or histograms.

4.) On page 17, please clarify whether the model fitted was multilevel and, if so, whether multilevel model fit was estimated using the wide format approach described in the present manuscript (e.g., as discussed by Mehta& Neale) or using the multilevel model fit indices reported by Mplus. The latter (e.g., RMSEA, CFI) have known difficulties detecting model misspecification at each level of analysis, particularly level 2.
The estimated model was a multilevel model. For clarification, we emphasized this on the first line of Paragraph 3.1: *"Note that with continuous data, the results as well as the fit statistics of the WF approach are identical to the results of the LF approach."* The RMSEA and CFI where you refer to are the results of our preliminary calculations to give an indication of model fit which assume that the data are continuous. For continuous data, the WF approach and the LF approach produce identical results. To make sure this is clear we added a sentence at the second last paragraph on P. 11: *"With continuous responses, the WF approach and the LF approach result in identical parameter estimates and standard errors."* The fit statistics are also identical. So, the difficulties with the RMSEA, CFI are not solved in this study.

5.) Please further develop the implications, or at least interpretation, of the illustration. Doing so will further demonstrate the utility of this statistical approach to advance education research.
Thank you for this very useful suggestion. We added a new section 'Conclusion' on P. 20 to discuss implications of the illustration: *"The STRS data shows that the WF approach can be used in datasets with a relatively small number of units in each cluster. Using the stepwise approach from section 1.3 and 1.4, various CFA models, including models with covariates and measurement bias restrictions, can be fitted to the STRS data. The DWLS in both the LF- and WF approch and the PML estimation method can estimate all suggested models. Due to the increasing dimensionality,*

the MML estimation method cannot fit models with error variances on the between-level (see Table 3). We hesitate to interpret the results of the the DWLS estimation method in the LF approach, as it shows a different pattern of parameter estimates compared to all other estimation methods."

6.) On page 19, I recommend that the authors clarify that the computational efficiency advantage of the wide format approach can decrease considerably as the number of observations grow.
Thank you for making this comment. We discuss this in the first paragraph of P. 21. *"As the maximum number of units increase, the size of the model syntax increases too. In the future, we plan to develop scripts that will automatically generate the model syntax."*

7.) Please add detail to figure captions so that each is fully encapsulated.
We have carefully reviewed all our captions, and added a few sentences where needed. Yet, we tried not to make them even longer as they are now.