We would again like to thank the reviewers for their helpful comments on the manuscript. As the study was reexamined and reworked, they served as very helpful guidelines. We have addressed each of these comments in the revised manuscript, as well as provided a point-by-point response below. Reviewer comments are in normal font and our responses are bold-faced.

1 Referee 1

1.1 Comments to the Author

- [Issues] that are specific to longitudinal studies (e.g., drop-out) should be discussed in light of the advantages of planned missing designs.

  This first iteration of this paper did not provide adequate context. This particular concern is addressed in the addition of an introductory paragraph.

- [There] is no justification for an inherent interest in a multilevel model. Why is this of greater interest than a latent growth curve or a simple repeated measures model?

  Our interest in studying a multi-level regression model stems from a lack of attention in recent literature. This is now addressed in the paper.

- The level of detail is more than readers need to get the essentials and pointing out that researchers have to choose how many blocks to include in any design is extraneous information and distracts from the larger argument.

  Much of this section was edited out and restructured for brevity and to avoid redundancies.

- Similarly, there is little mention about the mechanics behind multiple imputation - this is an important aspect of the study that deserves more attention than what is included. As a case in point, the authors mention that the choice of imputations is important but never mention why (e.g., improved efficiency). They also fail to mention the importance of the

imputation model.

We felt that as Sankhya B is an applied journal, some technical details could be omitted here. That said, this section has been rewritten with your comment in mind.

- [The] empirical example is so specific that I am not sure one can generalize beyond the limited set of simulations that were examined here. For example, can we reasonably assume that these findings would generalize to a continuous outcome? What about different parameters for the covariates? Are these sample sizes typical? And is there good reason to believe that a different design would have produced the same results? For example, is it reasonable to assume that there are no context or order effects in these questions that might play out in at least the split form designs? Of course, this is a flaw of any simulation; however, the highly specific nature of this study stands out. As an aside, why is there only one omitted response for day 5 (Q3)?

  Dr. Matthews??? As for only one response being omitted for day 5 the first iteration had an inaccurate representation of the designs.

- In table 7 it would be useful to have the true parameter values next to the parameter.

  This would be very useful, and has been added! DR. MATTHEWS CAN WE ADD THIS IS THE TABLES

- I was troubled by the seemingly arbitrary thresholds for what were unacceptable criteria. How was "Bias, Percent Bias, or MSE more than 40% greater than the complete-data value or a CI coverage of less than 90%" determined?

  DR. MATTHEWS???

- In section 4.3, the authors seem to confuse the idea of bias and efficiency.

  This mistake has been corrected, thank you!

2 Referee 2

2.1 Comments to the Author

- The authors compare their model with the one investigated by Rhemtulla et al. (2014), writing that "because our model of interest is a mixed effects model, though, we cannot be sure that this result will hold." It might be useful to spell out the differences between the two models, because a latent growth curve model can be specified as a mixed effects model. I believe there are three differences in the models: (1) we included a random slope, whereas yours is fixed; (2) we used a "latent basis" model in which the slope coefficients were estimated, whereas yours are fixed to specify linear change, and (3) you included observed predictors whereas we did not. It may be interesting to consider any differences in the results in light of these model differences, in particular, I would have predicted that a linear slope (and especially a fixed linear slope) is less affected by wave-level missingness than a random basis slope, which would seem to require more information at each time point in order to estimate.

  DR. MATTHEWS??? I can rewrite to address.

  model equation appears to be missing an error term in Level 1. If it isn't, that implies that data were generated with no prediction error - please clarify!

  The model equation has been rewritten to address this, and other concer

- I'm quite confused about the relation between the survey questions and the analysis model. Table 5 presents the survey questions, which I presumed to be the basis for the planned missing design, but the analysis model doesn't seem to include any of those items except "did you drink today" as a predictor. How is the split form design relevant to the variables included in the model? What was the missingness pattern of those items? More generally, I'd like to see the full model of the complete data that were generated and the details of all the missingness patterns that were imposed (i.e., which variables were missing data on which days in the 25%, 50%, and 75% missing conditions).

Our response goes here.

- In the split form design, is Day 5 / Q5 supposed to be missing? It's otherwise strange that Q5 is answered on 4/5 days and that Day 5 as 4/5 questions, as opposed to all the other days/questions being 3/5.

  Our response goes here.

- Further, please clarify whether that Table 6 presents the PM designs for a single participant. That is, I think another participant would be missing different questions on different days in the Split Form design, or missing different waves in the Wave Missingness design. Normally these design tables (like Tables 1-4) have a column for participants/forms, but Table 6 does not.

  Our response goes here.

- "The FMI is a measure of how much of the MI sampling variance comes from differences between imputations." That is how it is computed, but it is more generally a measure of the proportion of information that is lost as a result of missing data (e.g., see Savalei Rhemtulla, 2012 "On obtaining estimates of the fraction of missing information from FIML") "When determining if the PM designs are performing sufficiently...three possible PM designs." If different missingness structures are applied to the very same data sets, then the results are perfectly comparable. Further, it's not clear what "bias" and "average biases" refer to here - an estimate cannot be said to be biased based on a single data set, and it's not clear what's being averaged over.

  Our response goes here.

- "If a 95% CI has a true coverage of 90% then the Type I error rate is double what it should be." That's only true if the true parameter value is 0 and you're testing the hypothesis that the estimated parameter is different from 0. Otherwise these measures get at different things. (Indeed, there is no Type I error rate if the parameter value is not 0).

  Our response goes here.

- "While this fact was accounted for in the original simulations, it was

not in this set of simulations." This sentence makes it sound as though the second simulation design somehow failed to model the correlations among variables, rather than that low correlations were imposed. Am I right that the actual inter-item correlations are not presented anywhere in the paper? This is relevant information that is missing.

Our response goes here.

- Results are presented for the main effects in the model, but not for the variance of the random intercept - it would be interesting to see these results as well.

  Our response goes here.

- It is impossible to understand the results of the low inter-survey correlation condition without a clearer presentation of the design. If survey items were predictable from "baseline items" but not from each other, what were the correlations between these items and the baseline items? Were the baseline items used in imputation? And again, since the survey items don't seem to be in the model anywhere, how are these relationships expected to affect parameter estimation?

  Our response goes here.