We would like to thank the reviewers for their helpful comments on our manuscript. We have addressed each of these comments in our revised manuscript, as well as provided a point-by-point response below. Reviewer comments are in regular font and our responses are bold-faced.

We should also note that this revision is a little bit unusual in that the main author of the first submission has been dropped from this manuscript. Since then two other authors have been added to this manuscript, which has had parts largely re-written and the simulation study has been completely overhauled to conform to standards of reproducible research.

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.

  With any high frequency longitudinal study, there is a high probability of dropout. Using planned missingness reduces the burden on the participants, which often lowers the probability of drop out due to survey fatigue. It can also lower the cost of a longitudinal study. We believe this is addressed in the second sentence of the first paragraph of the new draft of the manuscript.

- [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?

  We chose to work with a multi-level model in this case because that was the analysis model of interest in Pellowski et. al. (2016). The data used in the aforementioned paper was used a basis for our simulation study, and so we wanted to continue to use the analysis model that was of interest in the original paper. The justification for using that model can be found in

Pellowski et. al. (2016).

- 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.

  We believe that we have addressed this issue by re-writing and cleaning up large sections of the manuscript, which we believe now contain the appropriate level of detail.

- 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.

  Complete case analysis is highly inefficient here because we would be removing nearly all of the observations. Multiple imputation gets around this and allows for more efficient estiamtion of the parameters.

- [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)?

  Certainly any simulation study suffers from the flaw that it is difficult to determined how generalizable the results are beyond the specific settings of the particular simulation study. However, the study design in Pellowski et. al. (2016) is a common design and the planned missingness designs studied here are

also common desgined.

That being said, we believe that results would be similar if the response was continuous and / or different parameters were chosen for the covariates. The sample size of this study is relatively small, but we still see reasonable results. In Pellowski et. al. (2016) the number of participants was 59. Sample sizes can be this low when attempting to collect data longitudinally at high frequencies such as daily (or even more often) due to the fact the participants need to be highly engaged in the study.

There should have been two omitted responses for Day 5 in the Split Form example in Table 4. This correction has been made.

How to respond to this: "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?"

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

  The true parameter values have been added to this table, which is now Table 5.

- 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?

  We have removed the mention of these cut-offs. However, they are based on suggestions put forth in Collins, L. M., Schafer, J. L., Kam, C.-M. (2001)

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

  That is correct. The mentions of efficiency have been removed from this section.

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.

  The complete data structure is best fit with mixed effects model and mimic the original data. We now exgend it to the scenario where we have missing by design. For the structure of this data, we believe this is the best model.

- The 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 was indeed written incorrectly. The model is now written as

  $$logit(p_{ij}) = \beta_{0j} + \beta_1 DrinkYN_i + \beta_2 ZAlcTox_i + \beta_3 Day$$

  where $\beta_{0j} = \gamma_0 + \alpha_j$ and $\alpha_j \sim N(0, \sigma^2_\alpha)$.

- 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).

The analysis model contains only one of the survey questions, but the other survey questions are used in the imputation model to impute the missing values of the question used in the analysis model. An example of the missingness pattern for low levels of missingness for split design can be seen in figure 1 in the left-most image.

- 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.

  We have removed this table from the manuscript and replaced with this Figure 1. Figure 1 shows an example of the missingness patterns for one single participant.

- 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.

  The table that the reviewer is referring to has been removed from the manuscript. It has been replaced by figure 1 and it is indeed an example of the missingness design for one single individual.

- "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.

We have changed this sentence to now read: "The FMI is a measure of the fraction of information that is lost due to missing data."

We have changed this sentence to read "In addition to comparing the results between each of the PM designs, the same analyses were also performed before missingness was imposed." We believe this addresses the reveiwers concern.

Finally, the bias and average bias are found by averaging across the different simulations.

- "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).

  Type I error rate refers to the rate at which the null hypothesis is rejected given that the null hypothesis is true. In this case, we are implicitly testing the null hypothesis that each regression coefficient is equal to the known value. (e.g. $\beta_1 = 0.28$). Since we know the true values of our population parameters, and we are generating simulated data sets based on those values our coverage rates are in fact estimating $1 - \alpha$, where α is the probability of making a type I error. There is always a type I error rate, even for parameter values different than 0. It is simply calculated by measuring how often the test is rejected (i.e. how often a confidence interval misses the true parameter value) conditional on the null hypothesis being true.

- "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.
The correlation matrix used needs to be added to the paper.

- 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.

  The main interest of this paper was the fixed effects coefficients, so we excluded analysis of the variance of the random intercept.

- 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.
  There was a major change to this section. We have added more information that hopefully satisfies the reviewer.