RE: #PMET-799, "Predictive Inference Using Latent Variables with Covariates"

Dear XXX:

We have received the reports on your manuscript, "Predictive Inference Using Latent Variables with Covariates" from an Associate Editor (AE) and four reviewers. The topic of your paper is certainly of interest to our readers, and while reviewer reactions were mixed, the overall view of the paper was positive. There are some areas that need attention. You have identified a potential problem, and you have shown some examples of inaccuracies that result from that problem. One question that comes up in these situations concerns how serious the problem is likely to be in practice (see Reviewers 3 and 4). Reviewer 4 notes that in some important respects, the conditioning models in the paper differ from those used in practice in testing applications. Reviewer 3 provides a reference to work that has looked at the practical consequences of the mismatch between conditioning and data models. You need to address this point in more detail. A second issue concerns the simulations.

Reviewers 1 and 4 each criticize the simulations as too limited to justify many conclusions, especially in relation to sample size. I would not want to have the paper switch the focus to simulation evidence, but some expansion of the simulations might help to bolster your message. Third, the imputation literature contains work on the issue of the mismatch between conditioning models. Much of this work is left out of the paper (see Reviewer 3's comments). Your paper makes a unique contribution by showing a particular mechanism leading to problems, but it is still worthwhile to cite the relevant literature. The reviewers have raised a number of other issues beyond these three, and you should address these as well.

On balance, I am going to reject this version of the paper, but I would strongly encourage you to revise the paper to address the issues raised by the reviewers. I think that all of these issues can be successfully addressed in a revision. When preparing your revised manuscript, please carefully consider the reviewer comments which are attached, and submit a list of responses to the comments. Your list of responses should be uploaded as a file in addition to your revised manuscript.

We look forward to receiving your revised manuscript.

Sincerely yours,

XXX

COMMENTS FOR THE AUTHOR:

Reviewer #1: While the paper presents an interesting idea it does so in a somewhat incomplete manner, and it makes the main point in a less rigorous formal argument than needed to be convincing. More specifically, on page 11 it is argued that "Intuitively, we see that when the prior for theta contains Y we have a circular argument. We want to determine the relationship between Y and theta using a theta that was determined given that we knew Y ." Note that this sentence only talks about the association of theta and Y, which is exactly what is argued before (p.8) is required, include all variables to get the associations right in the imputation model that provides a conditional distribution for theta. The issue can only be the estimation, and what the article states is just that "Elementary conditional density calculations force a shape on the conditional distribution of Y given theta and the Z variables" I am not sure whether this is simply a typo, or something else, but I do not think that the calculations force anything on the conditional distribution, I think what was meant is that the inclusion of Y in the conditioning model does something. What it does, however, is only insufficiently explained in section 3.1. It is neither proven rigorously, nor illustrated, it is more implied by verbal statements that the authors think they found evidence. I suggest to rethink the argument, and to sharpen it, and to provide a more rigorous treatment of what the expected effect of this circularity is if theta is an independent variable. Intuitively: If the imputation model for theta is correct, and we include all the things needed, then the conditional distribution of theta given the other variables, or any other of the other variables given theta and the remainder, should be correct, this is because if the model was indeed correct, we should be able to get very close to the joint distribution of theta, Z and Y (or U or W), so we should be able to also get the condtional distributions (in both directions) right.

The normal-normal example makes some assumptions that appear more consequential than the authors are willing to discuss. The fixing of beta_1 to 1.0 seems not to be as inconsequential as the authors try to make us think. While the scale of theta is arbitrary, the authors fix tau^2 (to a value > 0.0 - I assume - as it is customary, please clarify). This means that theta has a positive variance, so that with a fixed regression coefficient of 1.0 (= beta_1) we will see an effect on the conditional mean of Y (which does not need to be so, the regression could be unaffected by theta). So, there is not only loss in generality (if theta is assumed to have a positive variance tau^2) but also potential bias introduced in the estimates of the other coefficients by the assumption that theta enters the conditional mean

of Y with a regression weight of 1.0.

The simulation is too small-scale to be of much use: There is only one sample size condition, with N=280 (140 in focal and 140 in reference groups), and only two levels for the number of items. Also, it remains unclear how many replications were used, is there just one dataset simulated per condition? It almost seems like that (p19 bottom). The number of items varied in levels of 6 items (completely insuffient to measure anthing on a continuos scale, we can fit 3 located latent classes for 6 items and get fit comparable to IRT -see Lindsay Clogg & Grego, DeLeeuw & Verhelst, and others) and 50 items (where we have such a high reliability that conditioning is useless anyway) does make the whole simulation somewhat pointless. There is some indication that conditioning has a small bandwidth (not completely discrete measures, but also not reliability above 0.9, so in the range of 10 to 30 items we may see some gain). That is, in summary, the limitations of the simulation does not warrant the strong language, e.g., "the news is much worse" or "extremely bad news" on page 25 and elsewhere in the manuscript.

some random points:

p. 13: l. 39: There is an extra (unneeded) comma.

p. 17 equations (29) and (30) you may want to re-index, in other places the focal and reference groups are indexed 0 and 1, here they are 1 and 2.

p. 18 l.16. Novack should be Novick.

p. 19 l. 19. Refer to Birnbaum in Lord and Novick (1968) for the 2pl in place of or in addition to van der Linden and Hambleton (1997).

Reviewer #2: PM799
Review of Predictive Inference Using Latent Variables with Covariates

This is an excellent and well-written article that addresses a topic which is very important for secondary large-scale educational surveys such as PISA, PIAAC, TIMSS and PEARLS and NEAP. The results also pertain to secondary analyses in surveys in general.

Some minor adjustments need to be made, but I suggest to publish the article essentially in its current form.

Detailed comments.

Page 11, line 38. "to gives us" should be "to give us".

Page 11, line 34-52. Here I become a bit confused. First I was reading and expose about probability theory and now all-of-a-sudden P(Y,Z) becomes an empirical distribution which needs "enough data". Following that, this distribution (?) is integrated over. Keeping the expose in terms of probability is more adequate, I think.

Page 14, line 46. Sometimes Z has star, sometimes it has a tilde. It should be kept uniform.

Page 14, line 48. The subscript of tau-squared, that is Z*Y* is unnecessary and vanishes in the sequel.

Page 17, line 15. Substituting where?

Page 18, line 15. Novack should be Novick, and the reference is not yet in the reference list.

Page 18, line 40 and 49 (formula 34). Z should have a a subscript i.

Page 20, line 4. Discrimination parameters which are uniform on (0,2) is a bit strange, especially the lower end. It does not matter if the results are based on a number of replications. However, this is not made clear. Is there more than one replication with redrawn item parameters in each replication, however, this information is lacking.

Page 21, line 4-9. It is not just that standard errors of ability estimates increase with decreasing test length, also the bias increases, because Bayesian estimates shrink toward the prior when there is little other information.

Reviewer #3: Please see XXX.pdf, available on the Editorial Manager site.

Reviewer #4: See XXX.docx, available on the Editorial Manager site.