Responses to the Reviewers for Manuscript #PMET-799R1
"Predictive Inference Using Latent Variables with Covariates"

We appreciate the review comments. Our responses are in italics below.

Reviewer #1

1. My prior concerns were not addressed, to the contrary the meat of the paper was shortened to pages 18-21 ...

   The referee is correct that the meat of the paper is collected together for easy reference in section 4 (now pages 18-23) of the present manuscript, rather than scattered throughout the paper as in the original submission.

   ... which in addition still contain false assumptions. The conditional distribution of theta used to argue in section 4 and the theorems and corollaries therein are based on the posterior distribution of theta given Y and Z only (which is wrong).

   We apologize for lack of clarity in the text surrounding Theorem 4.1 and Corollary 4.3. The expressions $p(\theta|Y,\tilde{Z})$ [Theorem 4.1] and $p(\theta|U,\tilde{Z})$ do not represent posterior distributions at all, but rather conditioning models (as in equation 2 in the paper)—in Bayesian terms, these are prior distributions conditioned on covariates, not posteriors. The content of the theorem and corollary is that an institutionally specified conditioning model that involves Y severely constrains the selection of a research model $p(Y|\theta,\tilde{Z})$ by the secondary analyst; when the secondary analyst selects a research model that is inconsistent with the institutionally specified conditioning model, inferences based on institutional PVs are vulnerable to bias. We have reworded the text throughout Section 4, but specifically preceding Theorem 4.1 and Corollary 4.3, and following Corollary 4.3, to clarify these points.

   The dependency on item responses X is central here, but completely omitted from the central derivations that 'accuse' the use of PVs as independent as introducing 'wrong model' (a quite subjective term that sets up the reader's expectations rather than carrying any substance) bias. This wrong assumption made in theorem 4.1 and elsewhere in the section of course leads to a model that is completely determined, since all parts are, when P(theta,Y,Z) is known since then P(theta)=F(Y,Z) (which is not what is true for 'institional PVs').

Thank you for pointing out that the phrase "wrong model" was unnecessarily pejorative; we have removed it from the paper and simply made clear that we expect biases to occur when the SA's research model (equation 7 in the revised manuscript) is inconsistent with the primary analyst's conditioning model (equation 2 in the revised manuscript).

To summarize our response above, the central theme of Section 4 is exactly the mathematical relationship between the institutional conditioning model and the SAs research model when Y is included (directly as in Theorem 4.1 or by proxy as in Corollary 4.3) in the conditioning model, in the θ-independent case. When specification of the SAs research model is inconsistent with this mathematical relationship, inference using plausible values generated from a posterior depending on the conditioning model is vulnerable to bias.

The dependence of θ on X in the posterior is a separate issue, which we address in our next response (see below).

To the contrary, theta is mainly determined by X (the cognitive indicators - the variables most directly associated with theta - this association is much stronger than that between theta and Z, Y in typical cases), i.e. P(theta)=G(X,Y,Z) (almost = H(X), and the more items in X the more p(theta) = H(X), note that the NALS example also has a 3 dimensional theta = theta1,theta2,theta3, so the one dimensional theta PV used in the example is probably more determined by the items on the associated scale X1 and other 2 scales X2 and X3 than on any Y, and Z).

The referee appears to be citing the well-known property that, under commonly-made assumptions for a measurement model, the posterior distribution $p(\theta|X$, other fixed variables) depends only on X as test length grows (i.e., as measurement error decreases; see also our discussion at the bottom of p. 16 of the present manuscript). A consequence is that as test length grows, the magnitude of any bias due to mismatch between SA's research model and PA's conditioning model will decrease. We have added a sentence explicitly pointing this possibility out at the end of the paragraph following Corollary 4.3, on p. 22-23 of the revised manuscript.

2. The simulation study that made some faulty assumptions was not redone with correct assumptions, probably because this would have shown that the far reaching conclusions drawn from this simulation were unsubstantiated.

We agree with the referee that the simulations in the earlier version of the paper were unhelpful. We replaced the simulation study with an empirical example (Section 6) that provides evidence of the bias pointed to by the theory in Section 4.

3. The strong and almost dramatic language used in the first version that talked about how the Goldilocks approach has to be chosen, so that the dependent variable is 'never to be included' in the conditioning model if the PV is used as IV is completely gone from this revision. This is probably also because the initial derivations and logic had issues (as pointed out by prior reviews) that made this Goldilocks requirement unsubstantiated.

   We appreciate the referee's concerns with the dramatic Goldilocks language used in the first version of the paper; we removed it for that reason. Difficulties with including the dependent variable in the conditioning model, in the θ-independent case, are now discussed rigorously and dispassionately in Section 4 of the present manuscript.

   There were indeed some problems with the informal arguments in the very first submitted draft of the paper; the mathematical derivations in sections 3 and 4 of the present manuscript were developed, with the help of referees comments on our original submission, to correct these problems and provide a formal and rigorous basis for our argument about bias due to inconsistency between the SA's research model and the PA's conditioning model, in the θ-independent case.

4. The empirical example does not discuss potential alternative reasons for the differences found. i) the IRT model used in the example is different from the model used in the 'institutional' analysis. ii) the PVs generated for the NALS were likely based on a multidimensional conditioning model, iii) the MESE implementation needs to be checked against alternative implementation of similar models (using MPLUS, Latent Gold, or so).

   i) *We added a paragraph on page 30 of the revised manuscript, noting that the IRT model used in the MESE model is the same as the one used in the primary analysis and generation of PVs as per Kirsch et al, (2000). We set the item parameters to those reported by NCES in the MESE model since their estimates can be regarded as fixed and known, as discussed at the bottom of p. 7 of the present manuscript.*

   ii) *Our empirical example uses only the prose literacy domain. For the PV analysis, we use only the prose literacy PVs and for the MESE analysis we use only those items loading on the prose literacy domain, as described in the NALS user manual. We also note on page 30 of the revised manuscript that all NALS items load on only one scale, and while the conditioning model used in the production of the PVs is multivariate, separate estimates of the conditioning variables were estimated for each scale, so the additional items on the document and quantitative literacy scale can be seen as expanding the size of the conditioning model in the*

PV model for prose literacy.

iii) *Unfortunately, M-plus does not currently support a 3-PL IRT model, and for similar reasons it is not obvious how to build the full MESE model in LatentGold. We will happily provide R and WinBUGS code to the referees to check the correctness of our implementation, and will supply a web appendix for readers if the paper is accepted for publication.*

5. Finally, the results show that the PV based regression to a large extent agrees with the MESE model with covariates. It seems that the 2 MESE models (with and without covariates) show a much larger WRONG MODEL bias, namely if covariates are omitted we have the so-called omitted variables case, which leads to an overestimate of the race effect since the wrong MESE model (w.o. covariates) fails to include these relevant variables. Interestingly, the race effects are inflated, but not the effect of the PV skill variable for the wrong MESE model.

   Thank you for pointing out that specification 4 distracts from our point that the PV and MESE models provide different estimates for the coefficients in the regression equation that is the secondary analysts research model. We have omitted specification 4 and its estimates from the revised manuscript.

Reviewer #2

I verified that all comments by the reviewers were taken care of. I suggest to publish the article essentially in its current form.

Thank you for the positive assessment.

Reviewer #3

I have been out of deep contact with the psychometric world for years, and without doing more work than I have time to do now, the one suggestion that I can make for this paper is to try to build a bridge between the work on this topic with "plausible values" and the more general work on congeniality with multiple imputation. It might even be that some of the specific ideas proposed in the PV literature would have direct extensions to the MI world – I would think so.

I hope that this brief suggestion is helpful. In general, based on a cursory review of the paper and the reports that you already have, the work appears competent and relevant.

A central reference on the issue of "congeniality with multiple imputation" is provided by: Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. Statistical Science, 538-558.

Thanks for this reference. We refer to Meng's paper on page 4 in the introduction, and discuss how our work relates to his.