Point by Point Responses to Referees for PMET#799

Thanks for these helpful suggestions that have led to several improvements in our paper. We have changed the paper substantially since we first submitted it. Overall large changes include

- A focus on formalizing much of what was written in the earlier version of the paper. We have done that by adding Section 2, which outlines the various modeling components in educational survey analysis. Sections 3 and 4 then make our arguments separately about the case in which θ is a dependent variable in the secondary analyst's model and when θ is an independent variable respectively.
  In Section 3, we reformulate basic results for the θ-dependent case, going back to Mislevy (1991), in modern and fairly general notation. In Section 4, we have changed our argument subtly. We now argue that Y should be included in the conditioning set in order that the standard PV methodology produces unbiased estimates of s. While mathematically correct, this imposes a serious restriction on what the secondary analyst's model can be. As we show in Theorem 4.1, when Y is included in the conditioning set, it imposes a set of strict restrictions on the secondary analyst's model. Standard institutional PVs will only lead to unbiased results if the survey institution releases enough information that a secondary analyst could build a θ-independent model compatible with the conditioning model that generated the PVs. But to do so would likely lead to ethical dilemmas with confidentiality and privacy promises made when obtaining the data.
  We offer a suggestion of how to solve the problem we raise in the θ-independent case in Section 5 by using a "build-your-own" model and we suggest one in particular.

- We replaced the simulation study with an empirical example using data from the 1992 National Adult Literacy Survey in order to show the severity of the problem in practice. We show that we get very different results depending on the model used for analyzing θ. We believe the empirical example is a stronger example than the simulations because our simulations used only at most 2 covariates in the conditioning set. The empirical example uses PVs estimated by a very large conditioning set and compares them to results from the "build-your-own" model we suggest.

- We cut Section 4.1 from the paper for space constraints. Based on comments from Reviewer 4, we realized that the model in Section 4.1 is different from the models we discuss earlier in the paper. In the old Section 4.1, there is no primary or secondary model, just a model built from scratch. Thus the example does not help to make our point about the compatibility of the primary analyst's conditioning model and the secondary analyst's structural model.

We've responded to the specific suggestions of each reviewer below.

Reviewer 1 Comments

- "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 . . . 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."
  We have entirely rewritten what is now Section 4 (the θ-independent case) in a much more rigorous fashion. In addition, we have subtly changed our argument regarding the problem with using PVs as independent variables. We now argue through Theorem 4.1, that including Y in the conditioning set completely determines $p_{PA}(Y|\theta, Z)$. If a secondary analyst chooses a $p_{SA}(Y|\theta, Z)$ that is different from $p_{PA}(Y|\theta, Z)$, we argue that wrong-model bias will occur. Because of the complex nature of the conditioning model, a secondary analyst has essentially no chance of specifying a model consistent with the survey institution's modeling choices.

- "The normal-normal example makes some assumptions that appear . . ."
  We have cut the normal-normal example (the old Section 4.1) from the paper. We realized that the model in Section 4.1 is different from the models we discuss earlier in the paper. In the old Section 4.1, there is no primary or secondary model, but rather a model estimating θ and the regression coefficients simultaneously. Thus the example does not help to make our point about the compatibility of the primary analyst's conditioning model and the secondary analyst's structural model. We felt that the normal-normal example no longer added much exposition to an already long paper.

- "The simulation is too small scale..."
  In addition to cutting the normal-normal example, we also cut the simulations and instead replaced them with an example using data from the 1992 National Adult Literacy Survey (NALS). In our example, we analyzed the effect of literacy on log(weekly wages) for men ages 25-65 who work full time. We cut the simulations because we were concerned that the paper was becoming too lengthy as is and we felt that the more formalized arguments and the empirical example were sufficient. However, we would be happy to add the simulations back into the paper if others believed they provided additional support for our arguments.

- Extra comma on page 13
  This sentence is no longer in the paper.

- Page 17, equations (29) and (30)
  This example and equations (29) and (30) have been cut from the paper.

- Typo on Novick
  The Lord and Novick reference has been deleted from the paper so the typo is no longer there.

- Refer to Birnbaum
  We have made the paper more general rather than focus on the 2-pl IRT model and so the Lord and Novick reference is no longer necessary.

Reviewer 2 Comments

Reviewer 2's comments were fairly specific and detailed based on the earlier version of the paper. We have taken out the Normal-Normal example (which was previously on pages 14-18) and the simulations (which were previously on pages 18-24), many of the comments no longer apply.

The comment concerning page 11 lines 34-52 to keep the expose in terms of probability has been done. The new version of the paper formalized pages 11, lines 34-52 into Theorem 4.1 where we kept the arguments in terms of probability statements as suggested.

Reviewer 3 Comments

- As noted in the second paragraphs of Reviewer 3's comments, Y must be in the conditioning model when multiply imputing missing covariates.
  We now argue this in Corollary 4.1 (and for a specific case for Corollary 4.2) and their subsequent proofs on page 19.

- Paragraph 3 of Reviewer 3's comments note that issue arise when the conditioning model and the secondary analyst's model are not compatible.
  We also show, via Theorem 4.1 that including Y in the conditioning model can lead to wrong model bias. These comments were very helpful to us in building stronger arguments for our paper.

- "Recommendations for handling PVs"
  On pages 14 and 15, we have added more detail on the the standard procedure for using institutional PV's, described by Mislevy (1991, pp. 181-182).

- Arguments about inefficiency
  On page 18, we edited and improved our arguments about inefficiency in order to show how having a larger conditioning model can cause some inefficiency in the estimates of s.

- page 11, lines 17-19 priors in (2*)
  Our notation for conditioning models was rather confusing as noted by this comment of Reviewer 3's. Thus, in this version of the paper, we tried to improve on our notation by noting what models were from the primary analysts with a PA subscript and what models came from the secondary analyst with a SA subscript.

- "What is $Z \cap \tilde{Z}^C$
  We have cut that notation from the paper and instead argue on page 20 in Corollary 4.3 what happens when a U that is either independent or not independent of Y is in the conditioning model.

- Section 4.1 comments
  We have cut Section 4.1 from the paper for space constraints and replaced it with an empirical example using data from the 1992 National Adult Literacy Survey in Section 6. In addition, we realized that the model in Section 4.1, is a different from the models we discuss earlier in the paper. In the old Section 4.1, there is no primary or secondary model, just a model built from scratch. Thus the example does not help to make our point about the compatibility of the primary analyst's conditioning model and the secondary analyst's structural model.

- Exclusion restrictions
  Yes, we were using the concept of exclusion restrictions incorrectly and no longer use that terminology in the paper.

- Goldilocks rules in all situations
  We have added Section 6, which is an empirical example of the types of bias that can result from using institutional plausible values instead of a “build your own” model as we suggest in Section 5. We show that in some cases, (the coefficient on Black) the bias is quite large, whereas in other cases (the coefficient on Hispanic) the bias is a bit smaller. More work will clearly need to be done to effectively understand the extent of incompatibility problem. We show evidence that, in practice, the wrong-model bias can have substantial effects on the inferences one would draw.

Reviewer 4 Comments

- "Some real data from these assessment programs may be helpful to evaluate the bias..." We have added Section 6, which is an empirical example using data from the 1992 National Adult Literacy Survey which uses the PV methodology. We compare results from using the PVs to results that use two different conditioning sets (one "too small" and one "just right") from the Mixed Effects Structural Equations (MESE) model– a "build your own" model, we suggest in Section 5. We show that that bias ensues from using the PVs, although it is difficult to quantify (or even determine the direction of) the bias because the saturated conditioning set used in the PVs is so large.

- Goldilock's rules applying the multivariate latent variables
  In the new version of the paper, sections 2-4 are written as though θ is a univariate continuous variable. However, on page 10 in the footnote, we note that "X, Z and θ are extremely general multidimensional objects; they may have components that are continuous, discrete, etc. For ease of exposition, we will express all appropriate probability calculations as integrals, as if the variables involved were continuous. For other variable types, the integrals can be replaced with appropriate sums, Riemann-Stieltjes integrals, etc., as needed. The essential message of our work is the same." There is nothing in the arguments we make that require θ to be a unidimensional– although our example in Section 6 does focus on only one dimension of literacy.

- "Use of the term consistency"
  We have removed the term consistency from the introduction and conclusion for clarity and focused instead on discussing bias.

- Simulation study comments
  We have replaced the simulation study with an empirical example in Section 6 to show the types of bias that happen in real-world data. The example uses data from the 1992 National Adult Literacy Survey and it clarifies the extent of the problem of using PVs which the simulation study was not able to do with only 2 or 3 covariates. We cut the simulations because we were concerned that the paper was becoming too lengthy as is and we felt that the more formalized arguments and the empirical example were sufficient. However, we would be happy to add the simulations back into the paper if others believed they provided additional support for our arguments.

- Example 1 in Section 4.1 questions
  We have also edited out Section 4.1 because as pointed out here, this model is a bit different from the other models we discuss earlier in the paper. In Section 4.1, there is no primary or secondary model, just a model built from scratch. Thus the example does not help to make our point about the compatibility of the primary analyst's conditioning model and the secondary analyst's structural model.

- Minor comments: SEM as an abbreviation
  On page 23, SEM now appears after "structural equations model."

- Minor comments: θ as a response variable
  By this we had meant dependent variable, however, we can understand the confusion given that we had been discussing item response models. We have replaced the term response variable with dependent variable throughout the paper to help clarify.

- Minor comments: Page 8 lines 42-44
  We have expanded this sentence on pages 17-18 to make our arguments about inefficiency when the conditioning model is large clearer.

- Minor comments: page 13
  This section has been entirely rewritten to the new Section 4 and so these minor comments no longer apply.

- Minor comments: page 14
  This section has been cut from the paper and so these comments no longer apply.