RESPONSES TO REFEREE REPORT NUMBER 1

(*NOTE: The responses below come from the corresponding author of the manuscript. The other coauthors may choose to add their own views in subsequent posts.)*

1) **COMMENT:** ***"I believe the manuscript is an important contribution that significantly adds to our knowledge, and therefore can be published after minor revisions."***

RESPONSE: Thank you. The comments are very thoughtful and raise some issues which, to be honest with you, I had not previously considered.

2) **COMMENT:** ***"The authors follow the typical meta-analysis methodology, usually used outside economics. For example, the use of "fixed effects"' is not consistent with what an economist would imagine under such a term. One of the estimators is called ``weighted least squares,'' but it seems that virtually all estimates evaluated in the paper are weighted, which is puzzling. I worry that many readers will be confused by these terms. Perhaps this is a good opportunity to introduce terminology that would more intuitive and consistent. "***

RESPONSE: The referee is spot on about the confusing terminology. The original manuscript was written to connect with the set of researchers familiar with the respective estimators, and used the terminology that is common within that narrow slice of the literature. However, I agree that the manuscript would be improved by making it more accessible to those outside that literature. We will implement this suggestion in the next revision.

3) **COMMENT:** ***"In the present form the abstract of the paper will only be understandable to a few researchers narrowly specialized in the evaluation of the performance of various meta-analysis estimators. I am an empirical researcher who has written several meta-analyses in economics, but was confused by the terms used in the abstract."***

COMMENT: ***"The paper is written competently, but I would consider rewriting some parts, especially the abstract, which should provide us with the results of the study."***

RESPONSE: We will rewrite the abstract to make it more accessible.

4) **COMMENT:** ***"The performance of the estimators is evaluated under various settings, and the discussion of the numerical results is quite long; I worry that few readers will have the patience to go through all of this. Perhaps some parts could be moved to the appendix (especially the cases when only one estimate per study is***

assumed, which is not realistic in economics) and more space could be devoted to the interpretation of the results and implications for future use of meta-analysis."

COMMENT: ***"I think the authors should focus more on their contribution, which is the panel setting (each study reports several estimates). Almost all recent meta-analysis in economics work with several estimates from each study, so this is much more realistic and interesting than the rest of the analysis.***

RESPONSE: We will give thought to how to better focus the presentation of our results and the associated discussion. Certainly one option is to do what you suggest and move the single effect per study simulations into an appendix. However, these simulations -- which assume a simpler framework -- provide both a benchmark for the multiple effects per study simulations, and ease the exposition by allowing progressive generalizations of the DGP. At this point all I can say is that we will work on different ways of focussing our paper to address your concerns.

5) **COMMENT:** ***"The authors make several arbitrary choices in the design of their simulations; for example, see equation 3.B on page 5. Where do these numbers come from? I understand that arbitrary choices are inevitable, but would appreciate some justification of these parameters, and, potentially, evaluation how the results are robust to the choice of these parameters."***

RESPONSE: I fully appreciate the reviewer's point here. We have had numerous discussions amongst ourselves about this point. As the paper discusses, the trick is to find parameter values that simultaneously satisfy the following four criteria:

- Produce a realistic range of t-values for the estimated effects in the population of studies
- Produce realistic-looking funnel graphs
- Cause the percent of studies that "disappear" through publication bias to range between 0 and 80 percent.
- (where applicable) Produce realistic values of "effect heterogeneity"

We did this through experimentation and tried to present enough evidence that "it worked", without explaining (or, in some cases, being able to explain) why it worked.

At this point, I don't really know how best to address this comment. Here are my immediate thoughts:

- One reason for including the Stata programs in the Appendix is that the reader is free to try out their own parameter values. This seems to me the best way to address concerns about the parameter values that we use: Let the reader experiment with different values and see how it affects the results for themselves.
- At another level, it shouldn't matter if our parameter values are "arbitrary." At this point, the main message from our paper is:
  "*Wait! The evidence in favour of the PET and PEESE estimators/procedures is not nearly as compelling as one might think. We need to do some more experimentation before we declare any one or two procedures are 'the best'."*

In this sense, the fact that we can show that there exists at least one set of parameter values that produces realistic data sets but gets results different from the prevailing conventional wisdom is sufficient for our purpose.

To state it differently, our goal in this paper is relatively modest: To demonstrate that the question about "best practice for how to do a meta-analysis" is not settled. More work needs to be done.

6) **COMMENT:** ***The authors stress that the simulated meta-data should be as close as possible to the actual data samples used in meta-analyses. They could pursue this problem further in their analysis. For example, an important issue in meta-analysis is the unbalancedness of data: some studies report many more estimates than other studies. Should we weight estimates by the inverse of the number of estimates reported in each study to give each study the same weight, such as in Havranek and Irsova (2015)?"***

RESPONSE: The reviewer makes an excellent point. Like Havranek and Irsova (2015), I have also used an approach where, to check robustness, I weighted the effects by the number of estimates per study to give each study the same weight (http://pfr.sagepub.com/content/early/2015/02/10/1091142114568659). My short response is that this topic is one that lies outside the purview of the current paper but is definitely worthy of study.

My longer response is that this topic is related to the subject of clustering. Most meta-analysis studies cluster the effects from the same study in calculating standard errors. However, one can also use clustering to calculate GLS estimates of the effect, with "weights" related to the within-cluster correlation of effects. In the case of perfect correlation, this collapses to the 1 effect per study case. I suspect the advantages of giving each study the same weight depends on (i) the degree of correlation amongst effects from the same study, and (ii) the reason(s) for why studies differ in the number of effects they report.

7) **COMMENT:** ***"Another concern is precision-weighting with panel data. It is not clear how to interpret the results when weights (such as precision) are not constant within panels; for this reason, most Stata panel estimators do not allow precision weighting. Moreover, almost all economics meta-analyses include other moderator variables aside from the standard error. If the moderator variable is defined on the study level (the number of citations, for example), then precision weighting introduces artificial variation to that variable. In sum, I do not think we know precisely what it is that we estimate when we use precision weights for panel data, and the authors should consider non-weighted variants as well."***

RESPONSE: I don't think I understand the following comments by the reviewer:

- *"Most Stata panel estimators do not allow precision weighting."*

The paper presents panel versions of precision effect weighting in equations (11) and (13), and these are implemented in the do files for TABLES 8 and 9. I have also used,

in other settings, the xtgee function of Stata, which directly incorporates within-study correlation in calculating GLS estimates of the coefficients.

- *"If the moderator variable is defined on the study level (the number of citations, for example), then precision weighting introduces artificial variation to that variable."*

To be honest, I don't see the problem. One could just as well say this about the constant term, which has "artificial variation" induced into it when it is divided by the standard error. Perhaps it would be better to think of the transformation as weighting *observations* rather than weighting variables. Those observations with smaller standard errors are, so the theory tells us, more reliable than observations with larger standard errors, and thus should be given greater weight in the calculation of a mean effect. The logic is illustrated by comparing equation (5) with equation (4) in the paper. The underyling principle is the same when one includes moderator variables, as these explain how the "mean" changes systematically according to the moderator variables.

8) **COMMENT:** ***"Another crucial issue in meta-analysis is the potential (and likely) endogeneity of the standard error. When a method choice influences both the point estimates and standard errors in the same direction (think of using instrumental variables, for example), then SE is not exogenous to the point estimate. This is what I often observe in the data: tiny estimates are accompanied by tiny standard errors for no obvious reason other than methodology (but the precise causes are difficult to code). Such endogeneity leads to false detection of publication bias, so meta-analysts should instrument the standard error, preferably with something uncorrelated with method choices; for example, a function of the number of observations used in the original study (as in Havranek, 2015). If the standard error is not exogenous to the point estimate, then precision weighting is problematic even with one estimate taken from each study."***

RESPONSE: This is a great point which I have never seen raised before. As above, my short response is that this topic is one that lies outside the purview of the current paper but is definitely worthy of study. My longer response is that I definitely see this being a problem with IV estimation, which (ideally) corrects endogeneity bias at the cost of having a larger coefficient standard error.

The reviewer's comment relates to another point I have often wondered about: One relatively recent development in econometrics is a sophisticated set of "robust" standard error estimators that correct for (i) heteroskedasticity, (ii) serial correlation, (iii) cross-sectional correlation, etc. These "robust" estimators often, but not always, lead to larger standard error estimates compared to OLS-type estimators. Thus, we have the case where the more modern, and presumably more accurate, effect estimates may have standard errors that are larger than the dated effect estimates that use less sophisticated, OLS-type standard error estimates. Following the logic of "precision weighting", these more recent estimates would receive less weight.

It seems to me that the first step would be to assess whether this issue is practically important. One possible way to do this would be to undertake an SUR-

type “meta-regression” where there were two equations, one with dependent variable equal to the estimated effect size, and the other with dependent variable equal to the standard error of the estimate, and then test if the same moderator variable was jointly significant in both equations.

9) **COMMENT:** ***“Moreover, precision-weighted results are extremely sensitive to outliers. In each meta-analysis there are several estimates accompanied by implausibly large values of precision. What should we do with these values? Winsorize them as in Havranek et al. (2015)?”***

RESPONSE: Of course, the irony here is that the theory says it is precisely those outliers which can give the “best” effect estimates. This is illustrated in Figure 1 of our paper and and the associated discussion. See also: Stanley, T.D., Jarrell, S. B.and Doucouliagos, H(C). 2010. *Could it be better to discard 90% of the data? A statistical paradox*. American Statistician 64:70-77a.

It all comes down to whether those “implausibly large values of precision” are accurately measuring precision. This topic is related to the previous comments about standard errors. My own view is that, to date, standard errors have been taken at face value as accurate measures of estimator precision. Perhaps it is time for researchers to examine this assumption more closely.

10) **COMMENT:** ***“While I know the present study cannot answer all of these questions, they could perhaps resonate in the manuscript, and I would like to know the authors' perspective on these issues.”***

RESPONSE: While I am not 100 percent certain about how to handle this, I think mentioning the points above in a “Possible Directions for Future Research” section would address the reviewer’s concerns, and improve the manuscript.

FINAL THOUGHTS: I thank the reviewer for some exceptionally thoughtful responses. I would be keen to receive any further feedback he/she would care to give in light of my responses above.