RESPONSES TO REFEREE REPORT NUMBER 2

1) **COMMENT:** ***"This paper is an important contribution to the literature. ... the extensive simulations are valuable and have the merit of showing sensitivities and relative performance of methods that so far have been neglected in the literature."***

RESPONSE: Thank you. The reviewer's report makes a number of excellent points. We are confident that we can satisfactorily address them in the revision to our paper.

2) **COMMENT:** ***"Using Monte Carlo simulation it analyses the comparative strengths and weaknesses of standard meta-analytical procedures to deal with bias against statistically insignificant and wrongly signed estimates. The authors label this as publication bias, but I am not convinced that they cover all aspects of bias, or of publication bias specifically. In particular bias may occur due to intrinsic motivation and bias of the researcher (see, e.g., Doucouliagos and Paldam 2009 on development aid) and this is quite different from for example an elasticity where a priori the sign is known. So the bias studied in this paper is more limited and that needs acknowledgement somewhere in the paper (for example, p. 3 line 4)."***

RESPONSE: Following the reviewer's comment, we will revise the manuscript to make it clear that we do not consider all kinds of bias. However, as far as we know, the two types of biases that we study are the most commonly mentioned ones in the literature.

3) **COMMENT:** ***"The authors are not consistent in their treatment of 'wrong studies' in the first part of the paper the probability of such studies being published is set at 10% (without further motivation). In a latter part of the paper it is set a 0% (p.22 also without motivation)."***

COMMENT: ***"In order to reach a broader audience it would be helpful to provide clearer intuition regarding the different modelling strategies (explain better why the equations differ. In 3'.B you add 0.3. Help the reader to understand why this is sensible."***

RESPONSE: With respect to the first comment, the reviewer is correct that we use a different standard for deleting studies through "publication bias" for the panel random effects case. However, it is not 0%. On pages 21f, we state: "*In the case of bias against statistical insignificance, we assume that in order to be published, a study must have most of its estimates (7 out of 10, or more) be statistically significant. If the study meets that selection criterion, all the estimates from that study are "published." If the study does not meet that criterion, none of the estimates from that study are published. An identical "7 out of 10, or more" rule applies to publication bias against wrong-signed estimates.*"

With respect to both comments, one of the criticisms of Monte Carlo analysis is that the results are dependent on the particular parameters chosen. As noted in the paper, the trick is to find parameter values that simultaneously satisfy the following four criteria:

1) Produce a realistic range of t-values for the estimated effects in the population of studies
2) Produce realistic-looking funnel graphs
3) Cause the percent of studies that "disappear" through publication bias to range between 0 and 80 percent.
4) (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.

Here are some responses that we hope will satisfy the reviewer's concerns:

- 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 us to be 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 seem arbitrary. This is because the main message from our paper is: The evidence in favour of the conventional estimators/procedures is not nearly as compelling as one might think. Further theoretical developments and Monte Carlo experimentation where analytical results are not feasible should be undertaken before any one procedure is declared optimal in all common circumstances.

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.

4) **COMMENT:** ***"I am not convinced by the argument of the paper against the use of meta-analysis as a way to test hypotheses. It is quite counterintuitive that the meta-analyses can come up with a reliable point estimate but have such large failure rates when it is about the actual sign of that point estimate and the authors do not provide a convincing explanation. To me this looks like an error of reasoning."***

RESPONSE: We are not sure why this is counterintuitive. It is quite common to have estimators that are consistent in the coefficient estimates but inconsistent in their estimates of standard errors. To give a simple example, OLS with nonspherical errors produces consistent estimates of variable coefficients but produces inconsistent estimates of standard errors, leading to unreliable hypothesis testing. The reason for the incorrect estimates of the standard error in our Monte Carlo analysis is that none of the estimators appropriately model the DGP under publication bias.

5) **COMMENT:** ***"Moreover, all that the paper does is analyze this for a subset of methodologies ignoring for example 'more truly' hypotheses testing meta analyses such as Lazaronni and van Bergeijk (2014)."***

RESPONSE: We have read the Lazaronni and van Bergeijk (2014). The paper makes some nice methodological innovations, such as the use of factor analysis and generalized ordered probit. While we are not sure what the reviewer means by "more truly hypothesis testing.", we will investigate the relevance of their contribution for our paper.

6) **COMMENT:** ***"p. 21 it strikes me as very odd that a 55 per cent improvement is labelled as a qualitatively unaffected result."***

RESPONSE: The comment refers to Footnote #6 which states: "*For example, the Type I error rates in TABLE 8 for the WLS estimator fall from approximately 80 percent to approximately 25 percent when standard errors are clustered. While this is a large decrease, it does not qualitatively change the conclusions we draw from these experiments*." We acknowledge that the footnote should have been worded better. It will be revised accordingly. However, most researchers would still consider a 25 percent Type I error rate for a 5 percent significance level to be unacceptable.

7) **COMMENT:** ***"Why does the abstract contain acronyms? The last sentence of the abstract is awkward. The paper makes important contributions (listed on p.26); rearrange the abstract to flesh out these contributions."***

RESPONSE: The other reviewer also pointed out that we need to get rid of acronyms and overly technical jargon. We agree and will revise the abstract accordingly in the next revision.

8) **COMMENT:** ***"...add [in the abstract] "We set out a practical four step procedure that should be followed in meta-analysis."***

RESPONSE: Actually, we do not advocate a four step procedure. Here is what we say in the conclusion:

"*More specifically, our results suggest caution in employing the FAT-PET-PEESE procedure for estimating effects in the presence of publication bias (cf. Stanley and Doucouliagos, 2012, pages78f.). This approach advocates that researchers follow the following four-step procedure:*

- *STEP ONE: Test for publication bias using the PET specification;*
- *STEP TWO: Test whether there is a nonzero mean effect using the PET specification;*

- *STEP THREE: If one fails to reject the null hypothesis of no effect in STEP TWO, conclude that there is no evidence of an empirical effect.*
- *STEP FOUR: If one rejects the null hypothesis of no effect in STEP TWO, estimate the size of the effect using the PEESE specification.*

The findings of unreliability of hypothesis testing and the lack of general superiority of any one MA estimation in our Monte Carlo simulations suggest that there is insufficient evidence at this time to limit meta-analytic statistical inference to this approach."

9) **COMMENT:** ***"The authors have a tendency to put important arguments in footnotes. Footnote 1, footnote 2 and 7 are clear examples of what needs to be in the main text. In a sense footnote 2 is relevant for the conclusions section as well."***

RESPONSE: The revised manuscript will figure out a way to incorporate these footnotes into the text.

10) **COMMENT:** ***"The claim that previous studies that focus on publication bias assume that studies only produces one estimate needs references."***

RESPONSE: This claim refers to the previously mentioned studies (i.e., Moreno et. al, 2009; Stanley and Doucouliagos, 2014a, 2014b). The revision will make this clear.

11) **COMMENT:** ***"The authors (e.g. p. 9 refer to full percentages in the text) but percentages wt one extra decimal in the tables. Better to do this the same."***

COMMENT: ***"In the same vein on p. 16 numbers rather than percentages are given."***

COMMENT: ***"P.9 l.12. I would like to know the exact numbers/percentages for each of the two populations."***

RESPONSE: The revised manuscript will do a better job of making sure that the numbers in the text more closely match those in the tables. It will also be more consistent in using numbers/percentages.

12) **COMMENT:** ***"In the funnel plots it would be helpful to report the number of dots in the title of the plot (N=…) ."***

RESPONSE: The revised manuscript will report the number of dots/estimates in each of the plots.

13) **COMMENT:** ***"P 15 final but on line "to" generate."***

RESPONSE: The revised manuscript will change "so generate" to "to generate".

14) **COMMENT:** ***"P. 17 line 3 An."***

RESPONSE: The revised manuscript will change "a important difference" to "an important difference."

15) **COMMENT:** ***"P. 19 I would have liked to figures 4 and 5 at the beginning of the article. This is research reality. It provides the pictures that the article wants to simulate."***

RESPONSE: The revised manuscript will figure out a way of moving these figures closer to the beginning of the article.

16) **COMMENT:** ***"It would be good to have an overview table with the major conclusions (somewhere around page 25)."***

RESPONSE: The revised manuscript will include an overview table with major conclusions.

17) **COMMENT:** ***"P. 25 one but last line misses a "one"."***

RESPONSE: The revised manuscript will insert a "one" between "only" and "estimate per study."