Meta-analysis and publication bias: How well does the FAT-PET-PEESE procedure work?

A reply to Hong (*International Journal for Re-Views in Empirical Economics*, 2019)

W. Robert Reed*

International Journal for Re-Views in Empirical Economics, Volume 3, 2019-5, DOI: 10.18718/81781.14

JEL: B41, C15, C18
Keywords: Meta-Analysis, Publication Bias, Funnel Asymmetry Test (FAT), Precision Effect Estimate with Standard Error (PEESE), Monte Carlo Simulations, Replication Study

Data Availability: The R-code (Hong 2019b) to reproduce the results of Hong's replication (2019a) can be downloaded at IREE's data archive (DOI: 10.15456/iree.2018280.233725). The original programming by Alinaghi & Reed can be downloaded at Harvard's *Dataverse* (DOI: 10.7910/DVN/4IOLOP).

Please Cite As: Reed, W. Robert (2019). Meta-analysis and publication bias: How well does the FAT-PET-PEESE procedure work? A reply study to Hong (IREE, 2019). *International Journal for Re-Views in Empirical Economics*, Vol 3(2019-5). DOI: 10.18718/81781.14

1 Introduction

In an article in this journal (Hong, 2019a), Sanghyun Hong reports the results of replicating a recent paper by Nazila Alinaghi and myself (Alinaghi and Reed, 2018), henceforth AR. AR investigate the performance of the FAT-PET-PEESE (FPP) procedure. The FPP procedure is commonly used in meta-analyses for three purposes: (1) to test whether a sample of estimates suffers from publication bias (FAT), (2) to test whether the estimates indicate that the effect of interest is statistically different from zero (PET), and (3) to obtain an estimate of the mean true effect.

Hong's (2019a) replication focuses on a set of simulations that are built upon three data environments. In the first data environment, there is one true population effect. The only reason studies come to different estimates is due to sampling error ("Fixed Effects data environment – FE"). In the second data environment, there is a distribution of true population effects. Studies produce different estimates both because the underlying population effect is different, and because of sampling

*University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand. E-mail: bob.reed@canterbury.ac.nz

Received August 13, 2019; Accepted September 19, 2019; Published September 24, 2019.

error ("Random Effects data environment - RE"). Both the FE and RE data environments model a world in which each study only produces one estimated effect. The third data environment allows studies to produce multiple estimates, where the true effects are heterogeneous both within and across studies ("Panel Random Effects data environment – PRE").

Summarizing their results, AR (page 285) conclude that "the FPP procedure performs well in the basic but unrealistic environment of fixed effects... However, when we study its performance in more realistic data environments, where there is heterogeneity in the population effects across and within studies, the FPP procedure becomes unreliable for the first 2 purposes and is less efficient than other estimators when estimating overall mean effect. Further, hypothesis tests about the mean true effect are frequently unreliable."

Hong identifies two mistakes in AR. First, AR overstate the percent of estimated effects that are statistically significant in their simulated meta-analyses (cf. Hong's Table 2). Second, AR misspecify the FAT-PET-PEESE equation in their PRE simulations. This causes their simulations to underestimate rejection rates for the FAT and PET (cf. Hong's Table 3). This mistake also carries over to AR's performance comparison of the FPP procedure with two estimators that do not correct for publication bias (cf. Hong's Table 4).

After correcting AR's mistakes, Hong then extends their analysis. Unlike most other (all other?) studies of meta-analysis estimators, AR allow the simulated studies to produce more than one estimate. In particular, each study produces ten estimated effects in their simulations. While this is a step towards greater realism, it is still less than desired. Hong improves on this simulation environment by allowing studies to differ in the number of estimates they produce.

Hong summarizes the results of his replications as follows (page 14): "After correcting their mistakes and extending their analysis, I find that some values differ substantially, but the qualitative results remain the same." In other words, while individual results may differ, the poor performance of the FPP procedure in "realistic", artificial data environment is confirmed.

2 Background

Sanghyun Hong is a PhD student where I work (University of Canterbury). His replication analysis arose out of work with me on a follow-up research project investigating the performance of meta-analysis estimators (Hong and Reed, 2019). Part of that research reproduced simulation environments from AR. However, Hong was coding in R, while my code for AR was written in Stata. It was in the course of trying to reproduce the results in R that Hong discovered the mistakes.

When he first reported to me that he had found some mistakes in my programming code, he was embarrassed for me. When I suggested he write up the results and point out my mistakes, he resisted, thinking that would be disrespectful. I insisted that he do it and told him that this is how science needs to operate if it is to advance. And so he did, albeit reluctantly.

After IREE tweeted about publishing Hong's paper, Felix Schönbrodt (2019) a prominent psychological methods researcher, tweeted this in response:

> "Simulation studies need replications too: Independent implementations with the same settings (direct replications) increase the confidence in the specific result, broadening the parameter space (conceptual replication) tests generalizability and boundary conditions. An independent implementation of a simulation study of mine also discovered an error, which then could be corrected. ... let's not tie reputation too much to these bugs. Estimated industry average is 15 - 50 errors per 1000 lines of code; I expect more bugs in researcher's code. Bugs are normal; let's work together to find them!"

3 Why this is important

From my perspective, the major contribution of Hong (2019a) is not that he was able to confirm AR's findings. I would like to say that AR has had a major effect on the practice of meta-analysis. While the paper has only recently been published, it does not appear at this time to have received much notice.

The FPP procedure is well-entrenched in the meta-analysis literature. Thus, it is not surprising, and admittedly even appropriate, that researchers continue to use their familiar tool despite the results from a single study pointing out its weaknesses. In the first set of reports received by AR when they submitted their work to Research Synthesis Methods, one of the reviewers wrote: "The authors harshly criticize the FAT-PET-PEESE (FPP) technique due to Tom Stanley, which has been used by hundreds of meta-analyses (especially in economics, finance, and related fields)... Because the FPP method is so widely accepted and has been supported by many Monte Carlo simulations, the authors need to make a very strong case in order to warrant publication." The paper went through four revisions before it was finally accepted for publication. I suspect it will take many simulations confirming AR's results before researchers are willing to adopt alternative procedures.

The fundamental problem is that there are too few studies that compare the performance of meta-analysis estimators. Amongst the set of studies that use simulation experiments to assess performance, very few make their code available. For example, Stanley has published a large number of simulation studies that analyze meta-analysis estimators (Stanley, 2008; Stanley and Doucouliagos, 2014; Stanley, 2017; Stanley and Doucouliagos, 2017; Stanley, Doucouliagos, and Ioannidis, 2017). However, Stanley does not as a general practice make his programming code available. This makes it difficult for other researchers to build on his simulation results. Fortunately, it is now becoming more common for researchers to post their data and code (e.g., Carter et al., 2019). This allows other researchers to reproduce their testing environments to, for example, include additional estimators as they become available.

There are now a plethora of meta-analysis procedures available to researchers: Fixed Effects, Random Effects, WLS-Fixed Effects, WLS-Random Effects, Trim-and-Fill, Weighted average of the adequately powered studies, p-curve, p-uniform, PET-PEESE, Three-parameter selection model (3PSM), Andrews and Kasy's symmetric and asymmetric estimators, the endogenous kink estimator, and others. And there are a multitude of different types of simulation environments to test these estimators.

If meta-analysis research is to progress beyond adhoc adoption of disparate procedures, there needs to be a systematic approach of comparing estimators. A first step is that researchers should be able to reproduce the simulation environments of other studies to enhance comparability.

Therein lies the major contribution of Hong's work. Because he was able to identify mistakes in my programming code, I went back to the Harvard Dataverse archive where AR's code is posted, and alerted readers of my mistakes. I could then point them to Hong's code (Hong, 2019b), which IREE makes available via its website DOI: 10.15456/iree.2018280.233725. In this way researchers who want to reproduce AR's simulated data environments for use in their research can be assured that they are working with the correct code. I hope that Hong's work encourages other researchers to make their code publicly available, and that others will take it upon themselves to work through the code to check for mistakes. I see this as necessary if our understanding of "what works best" in meta-analysis is to advance.

References

Alinaghi, Nazila and W. Robert Reed (2018). "Meta-analysis and publication bias: How well does the FAT-PET-PEESE procedure work?" *Research Synthesis Methods* 9(2): 285–311. DOI: 10.1002/jrsm.1298.

Carter, Evan C., Felix D. Schönbrodt, Will M. Gervais and Joseph Hilgard (2019). "Correcting for bias in psychology: A comparison of meta-analytic methods." *Advances in Methods and Practices in Psychological Science* 2(2): 115-144. DOI: 10.1177/2515245919847196.

Hong, Sanghyun (2019a). " Meta-analysis and publication bias: How well does the FAT PET-PEESE procedure work? A replication study of Alinaghi & Reed (Research Synthesis Methods, 2018)." *International Journal for Re-Views in Empirical Economics* 3(2019-4). DOI: 10.18718/81781.13.

——— (2019b). " Meta-analysis and publication bias: How well does the FAT PET-PEESE procedure work? A replication study of Alinaghi & Reed (Research Synthesis Methods, 2018)." Dataset and Code. Version 1. *International Journal for Re-Views in Empirical Economics*. DOI: 10.15456/iree.2018280.233725.

Hong, Sanghyun and W. Robert Reed (2019). "A Performance Analysis of Some New Meta-Analysis Estimators Designed to Correct Publication Bias" *Working Paper* No. 19/04, Department of Economics and Finance, University of Canterbury.

Schönbrodt, Felix (2019, July 19th, 8:54 pm). "Simulation studies need replications too [Twitter post]." URL: https://twitter.com/nicebread303/status/1151414943192891392

Stanley, Tom D. (2008). "Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection." *Oxford Bulletin of Economics and statistics* 70(1): 103–127. DOI: 10.1111/j.1468-0084.2007.00487.x.

Stanley, Tom D. and Hristos Doucouliagos (2014). "Meta-regression approximations to reduce publication selection bias." *Research Synthesis Methods* 5(1): 60–78. DOI: 10.1002/jrsm.1095.

Stanley, Tom D. (2017). "Limitations of PET-PEESE and other meta-analysis methods." *Social Psychological and Personality Science* 8(5): 581–591. DOI: 10.1177/1948550617693062.

Stanley, Tom D. and Hristos Doucouliagos (2017). "Neither fixed nor random: Weighted least squares meta-regression." *Research Synthesis Methods* 8(1): 19–42. DOI: 10.1002/jrsm.1211.

Stanley, Tom D., Hristos Doucouliagos and John P. A. Ioannidis (2017). "Finding the power to reduce publication bias." *Statistics in Medicine* 36(10): 1580–1598. DOI: 10.1002/sim.7228.