Response by Tom Sanley to Data Colada [59], April 10th, 2017

Yes, invalid simulations based on unrealistic research scenarios do find shortcomings.

In a series of papers, we use meta-regression models to search for evidence of a genuine empirical effect when a research literature likely contains selective reporting (or publication bias or p-hacking) and also to reduce the bias already contained in this research area (Stanley and Doucouliagos, 2014; Stanley and Doucouliagos, 2017; Stanley, 2017; Stanley, Doucouliagos, and Ioannidis, 2017). Readers are referred to these papers, especially the last 2, for a detailed identification of the limitations of our methods in context and with suggestions about how their effects might be moderated. The cases where our methods perform poorly are caused by research literatures that contain little reliable scientific information, and in these same cases, conventional meta-analysis approaches are generally worse. We admit PET-PEESE is an aggressive approach to publication bias reduction that has more risk than more conservative weighted averages. For those who wish to take a more conservative approach, we offer the unrestricted WLS weighted average and WAAP, which is this WLS applied to only those studies identified to have adequate power (Stanley, Doucouliagos, and Ioannidis, 2017). These weighted averages are as good as random-effects when there is no selective reporting bias (or p-hacking), but consistently less biased when some studies engage in publication selection or p-hacking. In application, we consider our results to be only those findings that are robust to any reasonable and likely valid variation in methods, meta-regression models and approaches (Stanley and Doucouliagos, 2012).

Yes, invalid simulations based on 'this crazy assumption' under very unrealistic research conditions do show that PET is underpowered and that PET-PEESE can be biased downward when one further assumes, as Uri does, that all psychological research is without informative content. That is, when all psychological research is highly underpowered. We, on the other hand, assume that there is typically some genuine information of scientific merit in psychology and that some studies have adequate power (Stanley, 2017). But it is not necessary to throw psychology under the bus, to identify a single weakness in PET-PEESE, we have already done that and more. Uri, who referred Stanley (2017), merely takes one of the weaknesses that we readily reveal out of context.

1. Invalid simulations: Uri's simulations do not begin by creating outcomes for control and experimental subjects and then carefully calculating means, their variances, effect sizes, and SEs (as our sim's and all actual psychological research studies do). {Although he might change that before he posts this response.} Rather he just pretends that a random draw from a non-central t-distribution can represent an effect size (Cohen's d) from a psychological experiment. However, effect sizes will only have a non-central t-dist if the estimated d^ is independent of its SE. That is, the statistic, d^/SEd^, will only have a t-dist if d^ and SEd^ are independent– see, any mathematical statistics text. Unfortunately, the formula for the variance (and SE) of d^ has a second term involving d^ squared. Thus, the numerator and dominator of what is typically called a 't-value' (d^/SEd^) will not have a t-distribution (non-central or otherwise). This statistic has the conventional t-value only if the true effect size is zero. And, Uri's simulations make PET-PEESE look rather good when he assumes that true effect size is zero. Thus, Uri's simulations

are demonstrably invalid. They fail to simulate psychological as practiced and as their statistics are calculated. Uri's simulation use the wrong distributions. Furthermore his simulations are based on what he calls, "this crazy assumption: r(n, d) = 0" (Colada #58). Again, one only gets a t-dist if the d^ and its SE are independent. As everyone knows, SEd^ depends on sample size, n. So, if n and d^ have a nonzero correlation, then d^ and SE cannot be statistically independent, and d^ will not have a non-central t-dist. Thus, it is Uri's simulations that are making 'this crazy assumption,' not ours. Also, nearly every psychological study makes the same 'crazy assumption' according to Uri's view. t-tests (or equivalents) that are routinely reported in psychological research are only valid and their p-values are only correct if "this crazy assumption: r(n, d) = 0" is true. Here too, we have a much higher opinion of psychological research than Uri.

2. Unrealistic research conditions: Uri's central attack concerns the case where no study selectively reports (or p-hacks) its findings. However, in Colada #58, Uri ridicules anyone who would even ask the question of whether there is or is not selective reporting or p-hacking in any area of psychological research, because he believe that there is always selective reporting or p-hacking. Thus, Uri believes that the case that he pins most of his argument upon is very unrealistic, which makes his argument misleading, at best.. In the other cases of Colada #59, Uri assumes that **all** published psychological research has been selectively reported or p-hacked. This too, is very unrealistic. I have never seen a case (out of many dozens) where all reported findings are statistically significant in the same direction. If there is even one insignificant finding reported, then there cannot be 100% selectively reporting or p-hacking.

3. Throwing psychology under the bus: In addition to all of the above, Uri assumes that every research study uses very small sample sizes: n={12; 50} in order for him to find a case where PET-PEESE performs poorly. Many books and dozens of papers and surveys in psychology have made the case that unless a study has adequate power (Cohen's 80% convention), then it will cause more harm than good—see Cohen (1977), for example, and the hundreds of citations to Cohen's repeated warning about power. APA (2010) and the Psychonomic Society (2012), emphasize the important of ensuring that your study is adequately power. Yet, Uri assumes that nearly all psychological research ignores professional guidelines and are highly underpowered. It is easy to show that these small sample sizes cause almost all studies in Uri's simulations to be highly underpowered. A few exceptions will occur at the highest effect sizes combined with the largest sample sizes in his ranges. Of course, these few exceptions will become much fewer still when he skews his distribution of sample sizes. Thus, for most of Uri's simulations he is assuming that all of psychological research is underpowered and hence 'more harmful than homeopathy' to use his words. If an area of research is as bad as Uri believes it to be (highly underpowered and with notable selective reporting biases), PET will have low power and will therefore rarely suggest that there is something here of scientific merit. But this is exactly as it should do. If an area of research is entirely compromised, it would be irresponsible for a meta-analysis method to claim that there is some genuine psychological effect in it. Our simulations do not assume that psychological research is nearly as bad as Uri believes. We based our preferred sample size distribution on the survey of social psychology by Fraley and Vazire (2014), where some studies will have adequate power. However, to disclosure where our methods break down, we report the worst-case scenario where all research is underpowered but

regard this to be a rare exception. We honestly hope that it is a rare exception. If not, it is psychology (not meta-analysis) that is stuffed, to redirect Michael Inzlicht's more colorful assessment (*Slate*, March 6 2016).

T.D. Stanley

References:

American Psychological Association (2010). *Manual of the American Psychological Association, 6th ed.* Washington, DC.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York:Academic Press.

Fraley R.C. &Vazire S. (2014). The n-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE* 9(10): e109019. doi:10.1371/journal.pone.0109019.

Stanley, T.D. and Hristos Doucouliagos, *Meta-Regression Analysis in Economics and Business*, Oxford: Routledge, 2012.

Stanley, T. D and Doucouliagos, C. (2014). "Meta-regression approximations to reduce publication selection bias," *Research Synthesis Methods 5* (2014), 60-78.

Stanley T. D. and Doucouliagos, H. (2017). Neither fixed nor random: Weighted least squares meta-regression analysis, *Research Synthesis Methods* **8**, 19-42.

Stanley, T.D. (2017). "Limitations of PET-PEESE and other meta-analysis methods." *Social Psychology and Personality Science*, 2017, [Epub ahead of print] DOI: https://doi.org/10.1177/1948550617693062.

Stanley, T.D., Doucouliagos, C. and Ioannidis, J.P.A. (2017), "Finding the Power to Reduce Publication Bias," *Statistics in Medicine*, DOI: 10.1002/sim.7228. [Epub ahead of print]