American Psychological Association

Psychological Methods

ISSN: 1082-989X

2019, Vol. 24, No. 5, 576–577
http://dx.doi.org/10.1037/met0000237

REPLY

Consistency of Effects Is Important in Replication: Rejoinder to Mathur and VanderWeele (2019)

Larry V. Hedges and Jacob M. Schauer
Northwestern University

In this rejoinder, we discuss Mathur and VanderWeele's response to our article, "Statistical Analyses for Studying Replication: Meta-Analytic Perspectives," which appears in this current issue. We attempt to clarify a point of confusion regarding the inclusion of an original study in an analysis of replication, and the potential impact of publication bias. We then discuss the methods used by Mathur and VanderWeele to conduct an alternative analysis of the Gambler's Fallacy example from our article. We highlight that there are some potential statistical and conceptual differences to their approach compared to what we propose in our article.

Keywords: replication, meta-analysis, experimental design

We enthusiastically agree with Mathur and VanderWeele that the definition of whether the results of a set of research studies replicate one another should be based on the underlying effect size *parameters* and that evaluation of replication should be based on an analysis of effect size *estimates*. Call the effect size parameters $\theta_1 \ldots, \theta_k$ and the corresponding estimates $T_1, \ldots, T_k$. However, we believe that they may have somewhat misinterpreted our proposal for the evaluation of replication. We did not intend to argue that one could *define* whether the result of Study 1 (θ_1) was replicated in Studies 2 through $k \geq 2$ by examining only the results of Studies 2 through k alone (that is, $\theta_2, \ldots, \theta_k$). That would be logically impossible. Nor did we intend to imply that any analysis of Studies 2 through k alone could determine whether the results of those studies are similar that of the first study. Therefore, we agree that any method of evaluating whether the results of Studies 2 through k agree sheds no light on whether any of those studies agree with the results of Study 1.

What we did say was that the methods we proposed were "suitable for situations in which these considerations [publication bias] should be minimal, such as evaluating studies that are part of designed programs of replication or other situations in which study protocols have been registered in advance" (Hedges & Schauer, 2019b, 558). It is arguable how often publication bias should be a substantial consideration in evaluating research results. One of us published some of the first research documenting how much publication selection could, in principle, inflate effect size estimates, so we are certainly mindful of the possible effects (Hedges, 1984). However, as an empirical matter, we note that the effect size estimates from the original studies in the Many Labs Project were *smaller* than the average estimates in the registered replications for about half of the 16 experiments they considered (Klein et al., 2014). If there had been pervasive publication bias, the original effect should have been consistently larger. As a general rule, we doubt that most psychologists adjust the observed effect size estimates downward to account for publication bias in the interpretation of every study that they read.

In the absence of good reason to suspect publication bias, we would apply the methods we suggest to the entire corpus of studies to evaluate whether their results were reasonably consistent. The same can be argued Mathur and VanderWeele's proposed methods; the original effect estimate could easily be incorporated in the $P_{>q}$ metric, though as we argue below, this analysis serves a different purpose than the Q tests discussed in our article.

Mathur and VanderWeele describe alternative methods for assessing replication. One of these, P_{orig}, is formally a test of the null hypothesis that the original effect size parameter θ_1 is consistent with the distribution of effect size parameters from the replications; concretely, H_0: $\theta_1 \sim \text{N}(\mu, \tau^2)$ where $\theta_2, \ldots, \theta_k \sim \text{N}(\mu, \tau^2)$. This analysis is related to the section in our article about comparing a single study to several replications and is consistent with the meta-analytic literature on outlier analyses (see Hedges & Olkin, 1985; Viechtbauer & Cheung, 2010). Whether one interprets P_{orig} as the p value it actually is, or as a continuous metric of evidence as Mathur and VanderWeele argue, this analysis will only be meaningful or conclusive if P_{orig} is reliably small (with high probability) when H_0 is false. The probability that P_{orig} is small when H_0 is false is akin to the power of the test that generates P_{orig}.

We suspect that even if the original study has reasonable power, the P_{orig} analysis is likely to have low power. To see this, consider a simple case with two studies that are direct replications (the same design and statistical analysis) as we assumed in our article. If $\theta_1 > 0$, we would want an analysis of replication to be well-

Larry V. Hedges, Department of Statistics, Northwestern University; Jacob M. Schauer, Institute for Policy Research, Northwestern University.

Correspondence concerning this article should be addressed to Jacob M. Schauer, Institute for Policy Research, Northwestern University, 2046 Sheridan Road, Evanston, IL 60208. E-mail: jms@u.northwestern.edu

powered to detect a difference in effect parameters large enough so that $\theta_2 < 0$ (a reversal of sign). However, Hedges and Schauer (2019a) show that the power of the uniformly most powerful (UMP) test to detect such a difference will be unacceptably low (less than 80%) unless the power of *each* study to detect a non-null effect as large as θ_1 is over 98%. If multiple replications are conducted, Hedges and Schauer (2019a) show that there are similar constraints on the UMP tests of H_0: $\theta_1 = \mu$. Because P_{orig} concerns the difference between θ_1 and a distribution centered around μ, it will be even less powerful than an analysis concerning the difference between θ_1 and μ itself.

The low power discussed above is not necessarily a limitation of methods. The results of Hedges and Schauer (2019a) show that when we privilege a single study in an analysis of replication, even the *most powerful* analysis methods will almost certainly be underpowered, and that it will often be impossible to overcome that no matter how many additional replications are conducted. Whether or not P_{orig} is the most powerful test, it will be subject to the same limitations. Thus, while we have no objection to conducting the P_{orig} analysis, we would urge caution in interpreting the results. Further, it will almost always be impossible to justify the design of an ensemble of replications based on the power of such analyses.

Mathur and VanderWeele also use a method that estimates the proportion of studies in which true effect θ exceeds some criterion θ_O, which they call $P_{>q}$. We agree that this is an interesting descriptive quantity, and note that it assesses a different idea of "replication" than the Q tests we propose. The Q tests concern the consistency of experimental results, while $P_{>\text{q}}$ focuses on whether all results are greater than some threshold. Setting aside the difficult technical issues involved in the $P_{>\text{q}}$ analysis, we see the difference between $P_{>\text{q}}$ and Q in terms of the logic of the analysis, and how they operationalize "replication." The implication for $P_{>\text{q}}$ seems to be that if a large proportion of studies have positive true effects (e.g., $\theta > \theta_O = 0$) this implies strong evidence that an effect of a manipulation is positive. However, just because there is strong evidence that an effect is positive does not mean that studies necessarily successfully replicated. To think about the logic involved, consider a set of five studies that are direct replications (keeping all conditions, including subject population, as similar as possible). Suppose the pattern of true effects was {0.1, 0.2, 5.0, 10.0, 20.0}. All of the effects are positive (so that $P_{>0} = 1.00$), but they vary by over two orders of magnitude. Even though all of the effects are positive, we do not believe that most scientists would regard the first and the last of these studies, in particular, as demonstrating perfect replication. At the very least, there is a profound lack of experimental control, which might dwarf the experimental effect.

It appears that this suggestion has the same weakness as reliance only on statistical significance testing to interpret results: It makes the interpretation reliant on a dichotomization of the distribution of effect sizes. One way to avoid the conceptual problem raised above is to characterize the proportion of true effects greater than some θ_L and smaller than some θ_U,

$$P\{\theta_L < \theta < \theta_U\},$$

again setting aside the technical difficulties in doing so. However, focusing on the similarity of the θ values makes the analysis conceptually very similar to the analysis we propose. We showed that the noncentrality parameter λ could be considered a parameter characterizing heterogeneity of effects and interpreted in terms of the expected value (average) of the difference between θ values $E\{|\theta_i - \theta_j|\}$ (Equation 5) or the maximum difference between any θ_i and θ_j (Equation 7).

As we point out, estimation of heterogeneity parameters, such as λ or the variance component τ^2 (the variance of the θ_i) are other complementary analyses that might be used to assess replicability in conjunction with either proposal. While we believe that analyses of replication that focus on the similarity of true effects are in the same spirit as what we have suggested. They could take many forms and deciding which are the most satisfactory should involve technical considerations (e.g., accuracy) as well as practical and interpretational considerations.

We agree with the Mathur and VanderWeele that the statistical analysis and hypotheses tested should be chosen based on what we hope to learn about replication. If we are interested only in whether effects exceed some threshold, a different analysis (such as $P_{>\text{q}}$) may be warranted than if are interested in consistency of results across studies. However, it may be important to recall that the replication crisis in medicine was spurred by empirical evidence that there were not just a few contradicted results, but that there were many more "*initially stronger* effects in highly cited clinical research" (Ioannidis, 2005). Thus, at least in medicine, consistency of results is an important scientific consideration in assessing replication.

References

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9,* 61–85. http://dx.doi.org/10.3102/10769986009001061

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York, NY: Academic Press.

Hedges, L. V., & Schauer, J. M. (2019a). More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics*. Advance online publication. http://dx.doi.org/10.3102/1076998619852953

Hedges, L. V., & Schauer, J. M. (2019b). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods, 24,* 557–570. http://dx.doi.org/10.1037/met0000189

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association, 294,* 218–228. http://dx.doi.org/10.1001/jama.294.2.218

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahnik, S., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology, 45,* 142–152. http://dx.doi.org/10.1027/1864-9335/a000178

Mathur, M. B., & VanderWeele, T. J. (2019). Challenges and suggestions for defining replication "success" when effects may be heterogeneous: Comment on Hedges and Schauer (2019). *Psychological Methods, 24,* 571–575. http://dx.doi.org/10.1037/met0000223

Viechtbauer, W., & Cheung, M. W. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1,* 112–125. http://dx.doi.org/10.1002/jrsm.11

Received June 4, 2019
Revision received June 24, 2019
Accepted July 16, 2019 ■