Unequal Burden of Proof on Replication Studies

Marjan Bakker, Marcel A. L. M. van Assen, Chris H. J. Hartgerink, Michèle B. Nuijten,
Hilde E. M. Augusteijn, Paulette, C. Flore, Jelte M. Wicherts and Robbie C. M. van Aert
Tilburg University

Author Note

Marjan Bakker, Marcel A. L. M. van Assen, Chris H. J. Hartgerink, Michèle B. Nuijten, Hilde E. M. Augusteijn, Paulette C. Flore, Jelte M. Wicherts, & Robbie C. M. Aert, Tilburg School of Social and Behavioral Sciences, Tilburg University.

Correspondence concerning this manuscript should be addressed to Marjan Bakker, Tilburg School of Social and Behavioral Sciences, Department of Methodology and Statistics, PO Box 90153, 5000 LE Tilburg, The Netherlands, telephone: +31 13 466 2964, M.Bakker_1@uvt.nl.

Unequal Burden of Proof on Replication Studies

Maxwell, Lau, and Howard (2015) questioned whether psychological research is suffering from a replication crisis by pointing out that recent failures to replicate may turn out to be artifacts of underpowered replication studies. We agree with Maxwell et al. that failed (nonsignificant) replications are not evidence of no effect, and that (single) replication studies should not be presented as such. We also agree with Maxwell et al. on the importance of combining results of multiple studies with a meta-analysis. However, their proposal of the organisation of psychological science will put an unequal burden of proof on replication studies compared to the original study, which is problematic for scientific progress. Maxwell et al. propose *only* to conclude H_0 if the entire confidence interval for an estimated effect lies within (-.1, .1) (p.492), which may be achieved with 80% probability by many small replication studies or by means of a replication study containing 1,714 participants per group (their Table 1). Because this requirement for a large *N* does not apply to original studies but only to replications, this scenario results in an inefficient scientific enterprise wherein false positives continue to exist and effect size estimates in meta-analyses will often be biased.

A cumulative science seeks to answer questions like (1) “Does an effect exist?” and (2) “How large is the effect?”, and also: (3) “How can we discover the truth efficiently?” We performed a simulation study to investigate the answers within three different scenarios. The three scenarios are *Current* (science as we believe it is currently conducted), *Maxwell* (based on Maxwell et al., 2015), and *Ideal*. *Ideal* is inspired by how large-scale replication initiatives are currently implemented in practice (Klein et al., 2014; Simons, Holcombe, & Spellman, 2014). For all scenarios we simulated studies in a two independent groups design and added these to a meta-analysis until the range of the 90% confidence interval (CI; see note 8 in Maxwell et al.) around the estimated effect size *d* was smaller than 0.2 for *Current* and *Ideal*, and until the entire confidence interval for the estimated effect falls in (-.1,.1) with 80% probability for *Maxwell*. *Current* is further characterized by (1) publication bias for small studies (5% of nonsignificant results is published; Van Assen, Van Aert & Wicherts, 2015), but not for large studies (100% published), (2) the use of both original and replication studies for effect size estimation, and (3) a conclusion that an effect size is larger than zero if the estimated effect size is significantly larger than zero ($\alpha = .05$). The only differences between *Maxwell* and *Current* are that the original study is not used for effect size estimation (2), and H_0 is only concluded if the upper bound of the 90% CI < 0.1 (3). *Ideal* has no publication bias for replications (1), discards the original study (because it was subject to publication bias; 2), and concludes H_1 if estimated effect size is significantly larger than zero as in *Current* (3).

The conditions for the simulation study were as follows. The population effect sizes were $d = 0$ (no effect), $d = 0.2$ (small), and $d = 0.5$ (medium). We used a fixed effect meta-analysis in the simulations because the underlying effects sizes were constant within conditions. The sample sizes were 25 participants per group for the original study, and the sample sizes for the replication studies were small (25 per group), large (1750 per group), or mixed (probability of 1/71 for a large replication and 70/71 for a small replication; see osf.io/a2zmk).

The results of the simulation study are presented in Table 1. With respect to the first question (existence of the effect), the *Maxwell* scenario performed worst when true effect size was 0; it never discarded H_1 in case of small replication studies (100%), where the false positive rate was still very high for *Current* (99.1%) and nominal for *Ideal* (5%). The false positive rates of *Maxwell* and *Current* got much less when some or all large replications were conducted. Note that for nonzero population effects (i.e., $d \geq 0.2$), almost all effects were detected (acceptance of $H_1 > .95$) under each scenario. With respect to the second question, we found unbiased effect size estimated in *Ideal* or when only large studies were included, because no publication bias is present in these situations. Similarly biased effect size estimates were found with *Current* and *Maxwell* when N was small or mixed, with larger bias for small replications and smaller true effect size. Finally (third question), efficiency, measured by total sample size in the meta-analysis, was higher for *Ideal*, with up to 10 and 33 times as many subjects needed for *Current* and *Maxwell*, respectively, whenever the true effect size was zero and only small replication studies were conducted.

Our results indicate that the proposition by Maxwell et al. (2015) is not much of an improvement compared to the current scientific practice (which includes publication bias), since the practice will be even more inefficient than the current situation, poorly equipped to correct false positive findings, and provides biased estimates of actual effect sizes. Our simulation study also showed the importance of minimizing publication bias; unbiased effect size estimates are obtained at very high efficiency (van Assen et al., 2014), which highlights the importance of preregistered and adequately powered replication studies (Nuijten et al., 2015).

Maxwell et al. (2015) raised an important issue about power of recent replication attempts, but in our view insufficiently appreciated that original studies are subject to publication bias and often involve underpowered designs for the (subtle) effects that are typical for psychology (Bakker, van Dijk, & Wicherts, 2012). As long as there is publication bias and researchers continue to be rewarded for novelty rather than replicability, we should not raise the bars for replications higher than we set bars for original studies. If we do that, false positives and inflated effects will abound in the literature and we will witness little correction and a continued waste of resources.

References

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7,* 543–554. doi: 10.1177/1745691612459060

Klein, R. A., Ratliff, K. A., Vianello, M., Jr., R. B. A., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology, 45,* 142–152. doi: 10.1027/1864-9335/a000178

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70,* 487. doi: 10.1037/a0039400

Nuijten, M. B., van Assen, M. A. L. M., Veldkamp, C. L. S., & Wicherts, J. M. (2015). The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Review of General Psychology, 19,* 172-182. doi: 10.1037/gpr0000034

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at Perspectives on Psychological Science. *Perspectives on Psychological Science, 9,* 552-555. doi: 10.1177/1745691614543974

Van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2014). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods, 20,* 293-309. doi: 10.1037/met0000025

Van Assen, M. A. L. M., van Aert, R. C. M., Nuijten, M. B., & Wicherts, J. M. (2014). Why publishing everything is more effective than selective publishing of statistically significant results. *PloS One, 9,* e84896. doi: 10.1371/journal.pone.0084896

Table 1

Simulation Results

	Scenario	N = 50	N = mixed	N = 3500
μ = 0	Current	%H1: 99.4	%H1: 28.2	%H1: 8
		0.299 (0.076)	0.042 (0.074)	0.008 (0.033)
		[10805.2 (2237.51)]	[6548.39 (2571.41)]	[3500 (0)]
	Maxwell	%H1: 100	%H1: 28.7	%H1: 9.7
		0.285 (0.043)	0.026 (0.04)	0 (0.034)
		[35874.9 (4018.42)]	[7429.22 (3514.21)]	[3500 (0)]
	Ideal	%H1: 5	%H1: 5.6	%H1: 5.4
		0 (0.06)	0 (0.055)	0 (0.034)
		[1100 (0)]	[1885.91 (1300.86)]	[3500 (0)]
μ = 0.2	Current	%H1: 100	%H1: 100	%H1: 100
		0.525 (0.052)	0.313 (0.135)	0.206 (0.033)
		[4868.33 (937.138)]	[5174.77 (1324.43)]	[3500 (0)]
	Maxwell	%H1: 100	%H1: 100	%H1: 100
		0.52 (0.03)	0.254 (0.056)	0.2 (0.034)
		[16211.4 (1717.05)]	[7085.72 (3195.23)]	[3500 (0)]
	Ideal	%H1: 95.6	%H1: 96.5	%H1: 100
		0.2 (0.06)	0.2 (0.054)	0.2 (0.034)
		[1100 (0)]	[1877.34 (1297.35)]	[3500 (0)]
μ = 0.5	Current	%H1: 100	%H1: 100	%H1: 100
		0.687 (0.042)	0.622 (0.089)	0.502 (0.033)
		[1839.42 (263.802)]	[2812.49 (1301.02)]	[3500 (0)]
	Maxwell	%H1: 100	%H1: 100	%H1: 100
		0.686 (0.023)	0.571 (0.066)	0.5 (0.034)
		[6123.45 (477.382)]	[5778.93 (1512.94)]	[3500 (0)]
	Ideal	%H1: 100	%H1: 100	%H1: 100
		0.501 (0.061)	0.501 (0.053)	0.5 (0.034)
		[1100 (0)]	[1875.43 (1293.08)]	[3500 (0)]

Note. Results per condition depict: %H1 percentage = reflecting how often null-hypothesis of no effect in meta-analysis was rejected for *Current* and *Ideal* and percentage reflecting how often upper bound of 90% confidence interval around meta-analytic estimate was larger than 0.1 for *Maxwell*; Average of fixed-effect meta-analytic estimates; () standard deviation of average meta-analytic estimates; [] average of the sum of studies' total sample size; [()] standard deviation of the sum of studies' total sample size.