

Failure to Replicate: Bad Science or A Consequence of Using the Significance Criterion in Editorial Decisions?

David Lane
Rice University

The Open Science Collaboration found that only 36% of findings from published psychology studies were successfully replicated. Moreover, the mean effect size in the replication studies was about half of the effect size found in the original studies. Is this due to bad science or a consequence of using the significance criterion in editorial decisions?

The simulation conducted by Lane and Dunlap (1978) suggests it may be the latter. In this simulation, we assumed that researchers conducted experiments in which there were two independent groups (experimental and control) with 20 subjects per group, a true difference between groups of 8 points, and a population within-group standard deviation of 16. The data were analyzed using independent groups *t* tests (with assumptions of independence, normality and homogeneity of variance being true). All research was submitted for publication but only results that were significant at the 0.05 level were published. The power of the *t* test in this situation is 0.34.

The main findings of the simulation study were (a) that if a very large-scale replication study had been conducted on the published research, differences would have been significant only 34% of the time and therefore only 34% of the findings would have been replicated and (b) the mean difference between conditions in the published articles was a highly inflated 18.28 compared to the true difference of 8.00 which would be the expected mean difference in the replication study.

These results are strikingly similar to those of the open collaboration study in which only 36% of the findings were replicated and the mean replication effect size was about half the effect size in the published articles. Therefore, the low replication rates and smaller effect sizes of the Open Science Collaboration are entirely expected based on the assumption that only studies with significant effects are likely to be published.

Lane, D. M. & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, 31, 107-112.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349 (6251) PMID: 26315443