2025 ■ Vol. 21 ■ no. 1

Replication, a Hallmark of Good Science: Unraveling the Factors That Predict Replication Success

Malcolm Grant [a] , Brent Snook [a] & Cathryn Button [a]

[a]Department of Psychology, Memorial University of Newfoundland

Abstract ■ Considerable discussion in recent years has focused on failures to replicate findings in the psychological literature. In a Monte Carlo simulation of the research process, we examined several characteristics of research studies that might predict replication success (i.e., when both studies show similar effect sizes), and when estimated effect sizes reflect true effects. In our simulation, a successful replication was most likely when the initial findings had already been replicated once by the original author and when measurement reliability was high. As expected, greater replication success was also associated with narrow confidence intervals around effect-size estimates. However, sample sizes (i.e., those typically found in experimental psychological research) contributed relatively little to replication success. The estimates of true effect sizes were more accurate, aligning closely with the values specified in the simulation, under the same conditions associated with replication success. We discuss our findings in terms of how changes in research practices might produce more reliable psychological research.

Keywords ■ Replication, replication crisis, simulation, reliability .

✉ bsnook@mun.ca

10.20982/tqmp.21.1.p024

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer

Introduction

There has been considerable discussion concerning the replicability of findings in the psychological literature. While some authors claim that successful replications are not at issue (Schmidt & Oh, 2016), many studies have failed to replicate even widely cited findings (e.g., Jarrett, 2016). For example, the (Open Science Collaboration, 2015) found that only 36% of attempted social and cognitive psychology replications were "successful". Although this result was based on a very restrictive definition of replication success (i.e., $p < .05$ with effects in the same direction), it nevertheless produced a lively exchange of views about the state of psychological science and the causes of the failure to replicate (Asendorpf et al., 2013; Cairo et al., 2020; Gilbert et al., 2016; Lilienfeld, 2017; LeBel & Paunonen, 2011; Motyl et al., 2017). Some of the proposed causes have been small sample sizes (Asendorpf et al., 2013), unreliable measures (LeBel & Paunonen, 2011; Lilienfeld & Strother, 2020; Loken & Gelman, 2017), institutional incentives to publish single studies and a lack of rewards for replication attempts (Lilienfeld, 2017). The current research used a Monte Carlo simulation to unravel some of the factors that are associated with replication success. An advantage of simulations is that, unlike the typical research situation, the true effect size is known. Thus, we could assess the role of research practices in producing replicable effects and determine how close those effects were to "reality."

Within the Null Hypothesis Statistical Testing (NHST) tradition, a replication attempt is considered successful only if both findings reach statistical significance in the same direction. One problem with this definition of replication success is that an attempted replication that narrowly fails to reach significance but yields an estimated effect size that is close to the original effect size would nevertheless be considered unsuccessful. Rather than treating replication success or failure as a binary variable, we created a continuous index (*Replication Discrepancy*) by simply calculating the absolute difference between the effect-size estimates for a pair of studies. Thus, an attempted replication may

be considered relatively successful when two studies yield similar effect-size estimates (i.e. when the absolute difference between them is small). We also constructed an index of *Estimation Error* by comparing effect-size estimates with the true values, following the same procedure we used to create the replication discrepancy index.

This paper reports the results of a large-scale computer simulation of the research process. The simulation allowed us to explore how changes in methodological practices that are under the direct control of researchers affect the likelihood of replication by an independent researcher, as indicated by the replication discrepancy index. We varied *sample size* to reflect the values typically seen in the published literature and varied *reliability* across its range. We also examined the degree of replication success by an independent researcher in cases where the original researcher had or had not been relatively successful in *replicating their own research*. Replication of one's own work can serve as a proxy for the many details of research design that must be carried out carefully with attention to sound design principles. Conceptually, self-replication can be viewed as reliability on a large scale. Thus, successful replication of one's own work should be associated with successful replication by an independent researcher. In addition, we examined two variables that authors do not directly control (estimated effect sizes and width of confidence intervals). We expected that large estimated effect sizes and narrow confidence intervals would be associated with greater replication success and more accurate estimates of true effects.

Method

The focus of our simulation and the iteration unit, repeated many times, is referred to as a research project. In each *research project*, we simulated a simple experiment (Study 1) where participants are assigned randomly in equal numbers to one of two experimental treatment groups, a dependent variable is measured and an effect size is calculated. We then simulated an identical experiment by the same researcher but with a new sample of participants (Study 2). Lastly, we simulated a third experiment in each project (Study 3) to assess the likelihood of an independent replication.

To simulate the studies in each *Research Project*, we followed the procedure described by LeBel and Paunonen (2011; also see Lipsey & Wilson, 1993). We used a three-way factorial design, with four levels of *Population Effect Size*, five levels of *Sample Size* (as typically seen in the literature), and five levels of *Measurement Reliability*. Each of the resulting 100 cells of this design was simulated 5000 times, resulting in 500,000 research projects. We describe the specification of each of the independent variable levels below.

We set the overall population mean (μ) at 100 and the population standard deviation (σ) at 10. We took the following procedural steps for each research project.

1. One of four population effect sizes was selected with equal frequency: $\delta = 0.0, 0.2, 0.5$, or 0.8 (Cohen, 1992).

2. Two normally distributed treatment populations were defined with standard deviations set at 10 (as in the overall population) and means calculated as follows:

$$\mu_1 = \mu + \delta\,\sigma/2$$
$$\mu_2 = \mu - \delta\,\sigma/2$$

In the literature, effect sizes are usually reported as unsigned quantities with the direction implied only by the context. However, as our interest in this paper concerns replication, it was important to retain the sign of the effect. In each research project, we estimated effect sizes by subtracting M_2 from M_1 in the numerator. Except for those cases where the value of δ in the population was set at zero, this almost always resulted in positive estimates of effect size.

3. Samples of participant true scores were randomly selected from the two treatment populations. Pairs of treatment groups were always equal in size, but the size varied with equal frequency over each research project: n per group = 10, 20, 30, 40, or 50.

4. One of five reliability levels for the dependent measure was selected with equal frequency: $\rho_{yy} = .1$ to $.9$ in increments of .2.

5. The amount of measurement error variance (σ^2_{error}) needed to yield the selected reliability level was calculated as follows (*cf.* LeBel & Paunonen, 2011):

$$\sigma^2_{error} = \sigma^2(1 - \rho_{yy})/\rho_{yy}$$

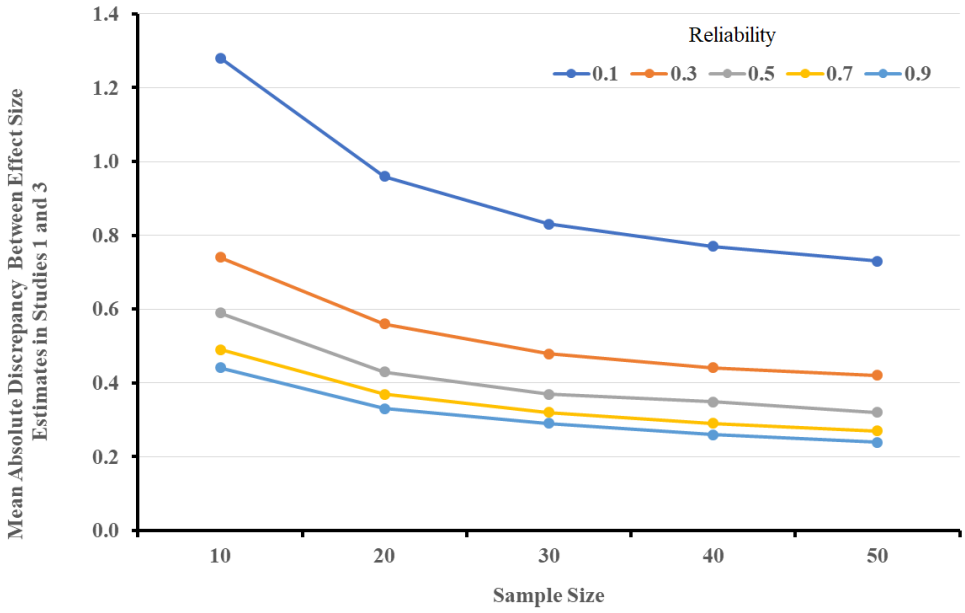
6. Each participant's measurement error score was selected randomly from a normal distribution with a mean of 0 and variance, as calculated in Step 5.

7. Participants' observed scores were calculated by adding their true scores (from Step 3) and error scores (from Step 6).

8. A second study, carried out by the original author, was simulated with the same true-score effect size, measurement error variance, and sample size as in the first study.

9. A third study, carried out by an independent researcher, was simulated with the same true-score effect size and measurement error variance as in the first two studies. In this case, however, a sample size of 50 was always used for both groups to increase the reliability of any observed effect size. We recognize that the sample size and other characteristics of Study 3 will affect the likelihood of replication. In this paper, however, we chose to adopt the perspective of the original author rather than the independent

Figure 1 ■ Mean self-replication discrepancy as a function of sample size and measurement reliability



researcher. Thus, we treated the results of Study 3 as simply an *indicator* of replicability and held the characteristics of this study constant in order to simplify interpretation.

10. Steps 1 to 9 were repeated 500,000 times.

For each study within a research project, we recorded the means (M_1 and M_2) and variances (S_1^2 and S_2^2) for the two groups and an estimate of the population variance (σ_{est}^2) obtained by pooling the sample variances. In addition, for each of the three studies, we calculated an estimate of the effect size corrected for unreliability (see Baugh, 2002; Bobko et al., 2001; Kanyongo et al., 2007), and the 95% confidence interval around that estimate using the approximation suggested by Nakagawa and Cuthill (2007). Lastly, we calculated the replication discrepancy index for Studies 1 and 2 and for Studies 1 and 3. We also calculated estimation error for Study 1.

Results

Using correlational and multiple regression analyses, we examined four variables to determine their usefulness in predicting replication discrepancy by an independent researcher and effect-size estimation error by the original researcher. In preliminary analyses, we included a fifth variable, the effect-size estimate. This variable, however, in later analyses, added no additional explained variance to predictions of either replication discrepancy or estimation error and was therefore dropped from further consideration. The four remaining predictor variables were the following: (1) *Sample Size*, (2) *Reliability*, (3) *Confidence Interval* around the effect-size estimate, and (4) the *absolute discrepancy* between the original author's effect-size estimates in Studies 1 and 2 (*Self-Replication*; see Cumming, 2008; Cumming & Finch, 2001; Schmidt & Oh, 2016). Descriptive statistics for the variables that were not manipulated are shown in Table 1, and the zero-order correlations among the variables are shown in Table 2. Given that only sample size and reliability were manipulated (and are under the control of a researcher), we further examined their role in predicting self and independent replication.

Self-Replication

The mean discrepancy in effect sizes for self-replicated studies (i.e., Studies 1 and 2), as a function of sample size and measurement reliability, is shown in Figure 1. As can be seen, for all levels of reliability we examined, there was a decrease in the discrepancy in effect size between the two studies as the sample size increased. Moreover, for sample sizes typically seen in psychological research (i.e., $n < 25$), there are substantial reductions in error by merely increasing reliability from .1 to .3. In addition, the discrepancy in effect sizes was smaller as the level of reliability increased. Studies with a sample size of 50 per condition and a reliability level of .9 produced the lowest levels of discrepancy in effect size across two studies conducted by the same re-

Figure 2 ■ Mean independent replication discrepancy as a function of sample size and measurement reliability

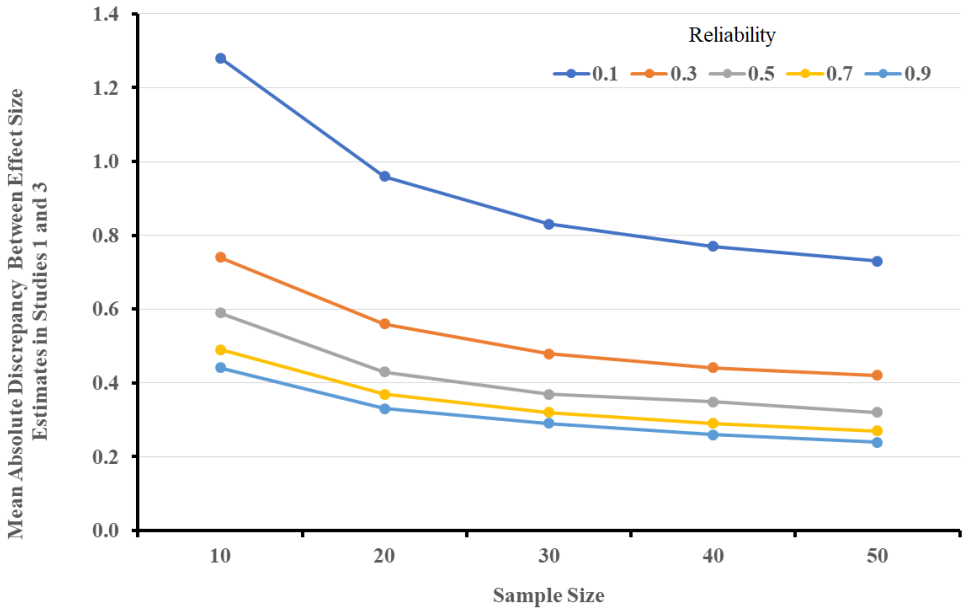


Table 1 ■ Descriptive Statistics for Variables in the Simulation

	Mean	SD	Minimum	Maximum
Effect Size (Study 1)	0.38	0.67	-7.11	8.21
CI Width (Study 1)	1.26	0.47	0.80	6.10
Self-Replication (Studies 1 & 2)	0.58	0.61	0.00	9.74
Independent Replication (Studies 1 & 3)	0.50	0.50	0.00	7.89
Estimation Error (Study 1)	0.41	0.43	0.00	7.71

Note. All statistics are based on 500,000 observations.

searcher.

Independent Replication

An examination of the correlations in Table 2 indicates that a relatively successful replication by an independent researcher was more likely when the original sample size was large, measurement reliability was high, when the confidence interval (CI) around the effect-size estimate was narrow, and when the discrepancy between the effect-size estimates in Studies 1 and 2 by the original author (*Self-Replication*) was relatively small. The mean discrepancy in effect sizes for independent replications (i.e., Studies 1 and 3), as a function of sample size and measurement reliability, is shown in Figure 2. As with self-replications, there was a decrease in the discrepancy in effect size between the two studies as the sample size increased; this pattern was observed for all levels of reliability. The discrepancy in effect sizes also decreased as the level of reliability increased. Again, for the sample sizes typically found in psychological studies ($n < .25$), substantial reductions in discrepancy are obtained when reliability is increased from .1 to .3 A multiple regression analysis revealed that the four predictor variables accounted for 36% of the variance in *Independent Replication*. To assess the predictors' relative importance, we examined each predictor's unique contribution when added to a model that included the other three predictors. The predictors, in order of their unique contributions, were as follows: *Self Replication* (7%), *Reliability* (5%), *Confidence Interval* (4%), and *Sample Size* (2%). The unstandardized regression equation was as follows:

$$\begin{aligned}\text{Independent Replication} = &-.50 \\ &+ .01 \times \text{Sample Size} - .44 \times \text{Reliability} \\ &+ .55 \times \text{CI} + .26 \times \text{Self Replication}\end{aligned}$$

Table 2 ■ Zero-order Correlations Between Variables in the Simulation.

	Effect Size (Study 1)	CI Width (Study 1)	Self Replication (Study 1)	Indep. Replication (Studies 1 and 2)	Estimated Error (Studies 1 and 3)
Sample Size (Study 1)	-.01	-.90	-.29	-.21	-.29
Reliability	.00	-.05	-.38	-.39	-.37
Effect Size (Study 1)		.10	.02	.03	.04
CI Width (Study 1)			.39	.34	.45
Self-Replication (Studies 1 & 2)				.51	.62
Independent Replication (Studies 1 & 3)					.77

Note. All correlations are based on 500,000 observations.

Effect-Size Estimation Error

The correlations in Table 2 show that effect-size estimates were more accurate when sample sizes were large, measurement reliability was high, the confidence interval around the effect-size estimate was narrow, and Studies 1 and 2 by the original author yielded relatively similar effect-size estimates. The mean estimation error as a function of sample size and measurement reliability is shown in Figure 3. As can be seen, for all levels of reliability, there was a decrease in mean estimation error as the sample size increased, and the estimated error was smaller as the level of reliability increased. As with discrepancy measures, absolute error was substantially reduced for studies employing the typical smaller sample sizes found in psychological studies, when reliability increased from just .1 to .3. Studies with a sample size of 50 per condition and a reliability level of .9 produced the most accurate estimates of the population effect size. A multiple regression analysis showed that the four predictors accounted for 50% of the variance in Estimation Error. The predictors, in order of their unique contributions, were as follows: *Self Replication* (12%), *Confidence Interval* (7%), *Reliability* (3%), and *Sample Size* (3%). The unstandardized regression equation was as follows:

$$\begin{aligned}\text{ESError} = &-.77 \\ &+ .01 \times \text{Sample Size} - .28 \times \text{Reliability} \\ &+ .60 \times \text{CI} + .30 \times \text{Self Replication}\end{aligned}$$

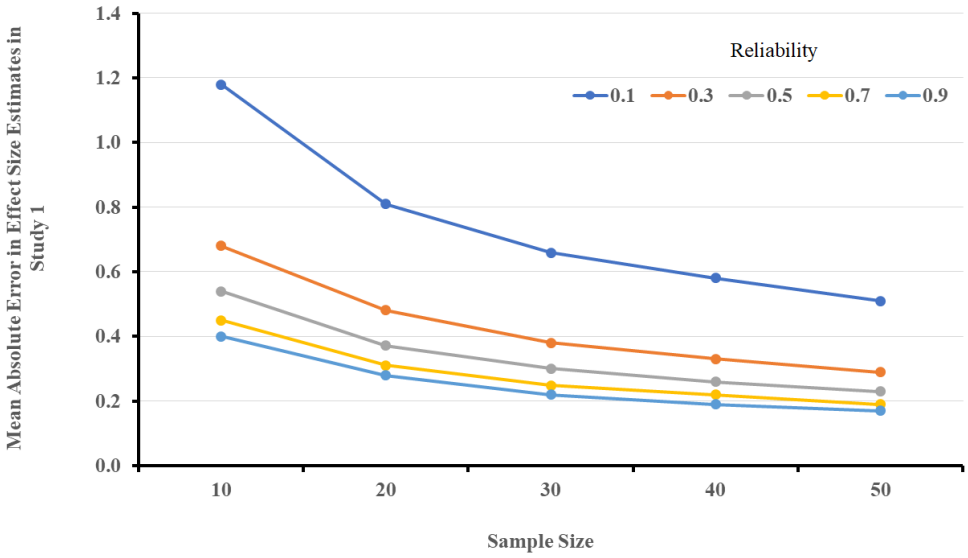
Discussion

In the present study, we first aimed to identify the methodological characteristics that predict the successful replication of research findings. We found that a successful replication by an independent researcher was more likely when the original researcher had successfully replicated their own work (i.e., lower replication discrepancy) and when the measures were reliable. To a lesser extent, we also found that the likelihood of replication success increased when the confidence intervals were narrow and sample sizes were large. Our second purpose was to identify the characteristics of studies that were associated with accurate estimates of the population effect (i.e., the actual effect size under investigation). Like replication success, the original author's replication discrepancy was the best predictor of population effect sizes, followed by confidence interval width, reliability, and sample size. Our findings support previous conclusions and recommendations that being able to produce similar effect sizes across two different studies is a good indicator that independent research will produce similar findings. Moreover, we found that the estimated effect sizes under the same conditions will be a good approximation of true effects in the real world.

Direct replications, while seen as a scientific necessity, are often held in low esteem and rarely published because of the widespread preference for novel scientific findings (see Martin & Clarke, 2017). However, the results of our simulation show that the most critical thing a researcher can do to increase replication success by an independent researcher is to replicate their own work. Our findings coincide with arguments from philosophers of science that direct replications are necessary because they help ensure that we can have confidence in findings (Koole & Lakens, 2012). It is worth emphasizing that replication success should be viewed as a matter of degree rather than strictly black and white (i.e., $p < .05$). Successfully replicating one's work - showing that two effect sizes are relatively close together - provides a check on sampling and measurement errors and situational influences (e.g., different participant samples). It also allows researchers to generalize results to different populations. In other words, direct replications enhance our confidence that others can reproduce the original effect and that the effect accurately reflects an underlying reality when valid measures are employed (Crandall & Sherman, 2016; Ioannidis, 2012; Schmidt, 2009). More broadly, a successful replication of one's research is a proxy for reliability on a larger scale.

Our simulation findings also support the truism that the reliability of measures is one of the crucial tenets of science; that is, the stability or consistency of measurement is

Figure 3 ■ Mean estimation error as a function of sample size and measurement reliability



essential to establish the validity (truth) of an inference and limit errors. If a measure produces inconsistent data, the researcher cannot determine what is being measured, and findings are less likely to be replicated. The value of having reliable measurement instruments is also part of our mundane experience. For instance, most would likely dispose of a bathroom scale if it gave us vastly different daily weights. Given the central importance of reliability and the numerous calls for researchers to attend to reliability, it is surprising that it is rarely mentioned in many published studies (Parsons et al., 2019). While we agree with Vazire et al. (2022) that replication is a shallow bar in this process, it is nonetheless an essential part of the research process. If the measure is inconsistent, there is no point in attempting to establish validity. While replicating one's work is not the ultimate end, it is a straightforward first step in developing good psychological measurements.

Our findings revealed that sample size, at least within the limited range we examined, had only a small impact on replication success. The sample sizes used in our research were deliberately chosen to reflect those commonly used in psychological studies. If researchers continue to rely on such small sample sizes, our simulation indicates that they must pay closer attention to the reliability of their measures and the replication of their own research before attempting to publish.

A limitation and direction for future research pertains to the ecological validity of our simulation parameters. The values we assigned to these parameters covered various situations encountered in psychological research and will be familiar to most researchers. We make no claim, however, that the parameter space explored here is necessarily representative of the actual parameter space that characterizes real-world psychological research. We have demonstrated that these variables have the *potential* to influence replicability in systematic ways. However, a more accurate evaluation of their influence will rely on simulations using parameter values that reflect their distributions in real-world research settings. We expect, for example, that the typical sample size will be larger with the proliferation of internet/online research, which employs much larger sample sizes than are typically used in laboratory studies. It will be interesting to see how simulating larger sample sizes will affect the role of the variables we studied here. Our findings highlight the importance of direct replications and underscore the need for reliable measurement techniques. In short, researchers need to prioritize reliability and conduct self-replications before publishing. Doing so may help to alleviate the problem of failed replications and lead to more accurate estimations of the effects under investigation.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., Wicherts, J. M., et al. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. doi: 10.1002/per.1919.

Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement*, *62*(2), 254–263. doi: 10.1177/0013164402062002004.

Bobko, P., Roth, P. L., & Bobko, C. P. (2001). Correcting the effect size of d for range restriction and unreliability. *Organizational Research Methods*, *4*(1), 46–61. doi: 10.1177/109442810141003.

Cairo, A. H., Green, J. D., Forsyth, D. R., Behler, A. M. C., & Raldiris, T. L. (2020). Gray (literature) matters: Evidence of selective hypothesis reporting in social psychological research. *Personality and Social Psychology Bulletin*, *46*(9), 1344–1362. doi: 10.1177/0146167220903896.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi: 10.1037//0033-2909.112.1.155.

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99. doi: 10.1016/j.jesp.2015.10.002.

Cumming, G. (2008). Replication and p intervals: P values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286–300. doi: 10.1111/j.1745-6924.2008.00079.x.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*(4), 532–574. doi: 10.1177/0013164401614002.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037–1037. doi: 10.1126/science.aad7243.

Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*(6), 645–654. doi: 10.1177/1745691612464056.

Jarrett, C. (2016, September 16). *Ten famous psychology findings that it's been difficult to replicate. british psychological society research digest*. Retrieved September 16, 2016, from https://digest.bps.org.uk/2016/09/16/ten-famous-psychology-findings-that-its-been-difficult-to-replicate/

Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, *6*(1), 81–90. doi: 10.22237/jmasm/1177992480.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*(6), 608–614. doi: 10.1177/1745691612462586.

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, *37*(4), 570–583. doi: 10.1177/0146167211400619.

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science*, *12*(4), 660–664. doi: 10.1177/1745691616687745.

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four scared cows. *Canadian Psychology*, *61*(4), 281–288. doi: 10.1037/cap0000236.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American psychologist*, *48*(12), 1181. doi: 10.1037/0003-066X.48.12.1181.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. doi: 10.1126/science.aal3618.

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? a snapshot of editorial practices. *Frontiers in Psychology*, *8*, 1–9. doi: 10.3389/fpsyg.2017.00523.

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., Skitka, L. J., et al. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of personality and social psychology*, *113*(1), 34–58. doi: 10.1037/pspa0000084.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Review*, *82*(4), 591–605. doi: 10.1111/j.1469-185X.2007.00027.x.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(62512), 943–950. doi: 10.1126/science.aac4716.

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive behavioural measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. doi: 10.1177/2515245919879695.

Schmidt, F. L., & Oh, I.-S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? or is it something else? *Archives of Scientific Psychology*, *4*(1), 32–37. doi: 10.1037/arc0000029.

Schmidt, S. (2009). Shall we really do it again? the powerful concept of replication is neglected in the social

sciences. *Review of General Psychology*, *13*(2), 90–100. doi: 10.1037/a0015108.

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, *31*(2), 162–168. doi: 10.1177/09637214211067779.

Citation

Grant, M., Snook, B., & Button, C. (2025). Replication, a hallmark of good science: Unraveling the factors that predict replication success. *The Quantitative Methods for Psychology*, *21*(1), 24–31. doi: 10.20982/tqmp.21.1.p024.

Received: 03/12/2024 ~ Accepted: 08/02/2025