Statistical (non)Significance ≠ (un)Successful Replication: The Importance of the Smallest Effect Size of Interest

Paul Riesthuis[1,2], Robert A. Cribbie[3] & Cristian Mesquida[4]

[1] Faculty of Law and Criminology, KU Leuven, Leuven, Belgium

[2] Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, the Netherlands

[3] Faculty of Health, York University, Toronto, Canada

[4] Human-Technology Interaction Group, Eindhoven University of Technology, Eindhoven, The Netherlands

Submitted to Meta-Psychology

Correspondence should be addressed to Paul Riesthuis: paul.riesthuis@kuleuven.be. The current manuscript has been supported by a FWO post-doctoral fellowship grant (1203824N).

Statistical (non)Significance ≠ (un)Successful Replication: The Importance of the Smallest Effect Size of Interest

Replications, direct or conceptual, are vital to estimate the boundary conditions or existence of empirical effects which is necessary for scientific progress. However, deciding which effects are worth replicating is difficult. The approach of Isager and colleagues (2024) facilitates this decision process by calculating an estimate of replication value derived from the citation count and sample size of original studies. Although this estimate of the replication value can be a good starting point to identify potential studies, it should not be a substitute for experts' judgment and substantive reasoning (Rainey, 2024). That is, the replication value of an effect depends on the specific field of study and needs to be contextualized. We argue that an important consideration for the value of replication studies is the decision of when researchers can decide whether a replication study is (un)successful. To be able to make such decisions, the smallest effect size of interest (SESOI; Lakens et al., 2018), also known as the minimally meaningful effect size (Beribisky et al., 2019), plays a central role (Camilleri et al., 2022).

SESOI and replications

The SESOI is the smallest effect that would lead researchers to conclude that a certain effect has practical or theoretical implications. Although basing the SESOI on theoretical predictions is a valid approach, it is often difficult as most psychological theories are verbal in nature and do not make quantitative predictions (Gruijters & Peters, 2020). Researchers should nonetheless identify what the smallest effect is that provides evidence for the theory and distinguish it, for example, from nuisance factors (e.g., experimenter effects, unblind research assistant; Wilson et al., 2021). Estimating the SESOI based on practical implications is more straightforward and takes into consideration another criteria for replication value which is societal impact (Isager et al., 2024). Either way, when a SESOI is established, researchers can conduct power analyses for their replication studies to make sure that the to-be-replicated study is sufficiently powered to examine whether the effect is truly meaningful and thus considered a successful replication (minimum-effect testing; Murphy & Myors, 1999) or whether the effect is too small to matter and thus can be considered as an unsuccessful replication (equivalence testing; Westlake, 1972)[1].

Ideally, original studies would set and justify a SESOI which the replication researchers can use or adapt due to different justifications (e.g., new insights, cost-benefit analysis; Camilleri et al., 2022). Then, a successful replication can be defined as one in which both the original study effect size and replication effect size and their 95%CIs exceed the SESOI. However, it is likely that original studies that are to be replicated did not set a SESOI because it only recently gained traction in psychological research areas such as sports and exercise (Mesquida et al., 2023) and eyewitness memory (Riesthuis et al., 2022). When a SESOI is absent in the original study, it is up to the replication researchers to specify a SESOI before the replication study is conducted and the replication can be regarded successful if at least the replication effect size and its 95%CI exceed the SESOI.

There is a myriad of ways in which the SESOI can be established, such as anchor-based methods, consensus studies, or cost-benefits analyses (Anvari & Lakens, 2021). Generally speaking, it is beneficial to focus on unstandardized effect sizes to grasp the practical implications (Baguley, 2009) and it can also facilitate setting the SESOI for theory testing. Important to note is that the estimation of the SESOI of replication studies does not need to be foolproof as long as it is justified appropriately (Riesthuis, 2024) and, for registered reports, preferably agreed upon by the reviewers. In fact, we argue that an interesting approach to increase the value of a replication study would be to have the replication researchers and the original authors together decide upon the SESOI which can

[1] We focus on replications for original studies that found statistically significant effects. Replications can also be conducted for null findings and then the criteria for when a replication is (un)successful is reversed.

then be used to examine when a replication is considered successful before it is conducted. This is because to design a study that allows for strong inferences (Platt, 1964), it is important to specify which results would lead researchers to claim whether a replication study has been (un)successful.

When the SESOI is not carefully decided upon, the added value of a replication study is unknown because it becomes unclear when a replication is successful or not. That is, researchers might implicitly set their SESOI by relying on Cohen’s benchmarks or field-specific effect size distributions in their power analyses. Both approaches are problematic because the former fails to take into account the context of a study (Cohen, 1988, p. 25) and the latter does not provide information on whether an effect is practically or theoretically meaningful and may be overestimated due to publication bias (Panzarella et al., 2019). Alternatively, a frequently used approach to determine the required sample size and subsequently the minimally detectable effect size is through the small telescope approach (Simonsohn, 2015). Again, this does not provide information on whether the effect is of any practical or theoretical interest and fails to take into account the context of the study (Primbs et al., 2022).

Without a carefully considered SESOI, researchers may rely on decontextualized methods such as statistical significance to determine whether a replication was successful or not (Schauer & Hedges, 2021). On the one hand, such reliance on statistical significance can lead to trivial effects, due to large sample sizes, being regarded as successful replications. On the other hand, it can lead researchers to conclude that the replication failed even though a practically or theoretically interesting effect exists but the sample size was too small to detect such an effect. However, when replication studies are conducted, researchers typically want to provide evidence that the effect is truly meaningful or too small to care about which means that minimum-effects or equivalence tests need to be conducted. Of course, a single replication study cannot rule out whether an effect does or does not exist (Nosek & Lakens, 2014), but replications should provide solid evidence in favor of one or the other. However, when only null-hypothesis significance tests are employed and contextualized SESOIs are lacking, costly replication studies might yield inconclusive evidence.

An illustrative example

To illustrate the importance of setting a SESOI for replication studies, we will discuss a recently published multilab replication concerning cognitive-dissonance theory (Vaidis et al., 2024). The authors provided thoughtful methodological and theoretical considerations of why they chose to replicate the induced-compliance paradigm of Croyle and Cooper (experiment 1, 1983). The replication value for this study would be 1.64[2] (Isager et al., 2024). However, for the power analyses, the authors relied on Cohen's benchmarks (i.e., .20 and .25 were used as small effect sizes) and meta-analyses to indicate their effect sizes of interest for the main hypotheses. Justifications for why these effect sizes were of interest, practically or theoretically, were lacking. Importantly, for this specific multilab replication study, 39 labs were involved and almost 5,000 participants were recruited.

The importance of setting a SESOI becomes apparent when examining their results, specifically when assessing the confidence intervals (CI) which can be used for minimum-effect and equivalence testing (Smiley et al., 2023). That is, when interpreting the 95%CIs it is observed that for the null findings the CIs are smaller than the Cohen's *d's* of .20 or .25 (e.g., findings on postessay attitude: $t(2{,}751.46) = -0.79$, $p = .79$, $d = -0.03$, 95%CI = [–0.10, 0.04], Vaidis et al., 2024, p. 15) which provides evidence that the results are equivalent based on the given SESOI. However, for the statistically significant findings (e.g., $t(2{,}408.92) = 6.51$, $p < .001$, $d = 0.26$, 95%CI = [0.18, 0.34], p. 15), the 95%CIs frequently included the

[2] $RV_{Cn} = \frac{375}{42} \times \frac{1}{\sqrt{30}} = 1.63$ (unweighted citation count). Using the theory of cognitive dissonance citation count (Festinger, 1957), RV_{Cn} = 36.5.

Cohen's *d* of .20 or .25 which indicates that the results are inconclusive (Riesthuis, 2024). Specifically, the effect might be meaningful but it could also be equivalent. However, more problematic is that these interpretations all hinge on the decontextualized SESOIs based on Cohen's benchmarks. In other words, it remains to be seen whether the replication can be deemed (un)successful when more thoughtful SESOIs are established and statistical significance is not solely relied upon to make this decision. As replications can be costly and time consuming, careful consideration of when a replication is (un)successful via establishing the SESOI should be contemplated when assessing the replication value.

Conclusion

In this commentary, we argue that to further estimate the replication value of specific studies, the SESOI should be determined. Specifically, we argue that the SESOI is a key component of replications because of its role in deciding when a replication is (un)successful. That is, when a SESOI is set, power analyses and data analyses for minimum-effect and equivalence tests can be conducted to establish whether the effect is practically or theoretically meaningful or not.

Conflict of Interest

All authors declare no conflict of interest.

Author Contributions

Paul Riesthuis: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing. **Robert A. Cribbie:** Writing – review & editing. **Cristian Mesquida:** Writing – review & editing.

References

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104159. https://doi.org/10.1016/j.jesp.2021.104159

Baguley, T. (2009). Standardized or simple effect size: What should be reported?. *British Journal of Psychology*, *100*, 603-617. https://doi.org/10.1348/000712608X377117

Beribisky, N., Davidson, H., & Cribbie, R. A. (2019). Exploring perceptions of meaningfulness in visual representations of bivariate relationships. *PeerJ*, *7*, e6853. https://doi.org/10.7717/peerj.6853

Camilleri, C., Beribisky, N., & Cribbie, R. (2022). The minimally meaningful effect size: A vital component of pre-registrations. https://doi.org/10.31234/osf.io/jbgtm

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Croyle, R. T., & Cooper, J. (1983). Dissonance arousal: physiological evidence. *Journal of Personality and Social Psychology*, *45*(4), 782- 791. https://doi.org/10.1037/0022-3514.45.4.782

Gruijters, S. L., & Peters, G. J. Y. (2020). Meaningful change definitions: Sample size planning for experimental intervention research. *Psychology & Health*, *37*, 1-16. https://doi.org/10.1080/08870446.2020.1841762

Isager, P. M., van 't Veer, A. E., & Lakens, D. (2024). Replication value as a function of citation impact and sample size. https://doi.org/10.31222/osf.io/knjea

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*, 259-269. https://doi.org/10.1177/2515245918770963

Mesquida, C., Murphy, J., Lakens, D., & Warne, J. (2023). Publication bias, statistical power and reporting practices in the Journal of Sports Sciences: potential barriers to replicability. *Journal of Sports Sciences*, *41*, 1507-1517. https://doi.org/10.1080/02640414.2023.2269357

Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology*, *84*(2), 234. https://doi.org/10.1037/0021-9010.84.2.234

Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*. https://doi.org/10.1027/1864-9335/a000192

Panzarella, E., Beribisky, N., & Cribbie, R. A. (2021). Denouncing the use of field-specific effect size distributions to inform magnitude. *PeerJ*, *9*, e11383. https://doi.org/10.7717/peerj.11383

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, *23*, 208. https://doi.org/10.1037/met0000126

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, *146*(3642), 347-353. https://doi.org/10.1126/science.146.3642.347

Primbs, M. A., Pennington, C. R., Lakens, D., Silan, M. A. A., Lieck, D. S., Forscher, P. S., ... & Westwood, S. J. (2023). Are small effects the indispensable foundation for a cumulative psychological science? A reply to Götz et al. (2022). *Perspectives on Psychological Science*, *18*(2), 508-512. https://doi.org/10.1177/17456916221100420

Rainey, C. (2024). Use and misuse of a fast approximation: Not a criticism, but a caution. https://doi.org/10.31222/osf.io/8z45v

Riesthuis, P., Mangiulli, I., Broers, N., & Otgaar, H. (2022). Expert opinions on the smallest effect size of interest in false memory research. *Applied Cognitive Psychology*, *36*, 203-215. https://doi.org/10.1002/acp.3911

Riesthuis, P. (2024). Simulation-based power analyses for the smallest effect size of interest: A confidence-interval approach for minimum-effect and equivalence testing. *Advances in Methods and Practices in Psychological Science*, *7*(2). https://doi.org/10.1177/25152459241240722

Schauer, J. M., & Hedges, L. V. (2021). Reconsidering statistical methods for assessing replication. *Psychological Methods*, *26*(1), 127. https://doi.org/10.1037/met0000302

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559-569. https://doi.org/10.1177/0956797614567341

Smiley, A. H., Glazier, J. J., & Shoda, Y. (2023). Null regions: a unified conceptual framework for statistical inference. *Royal Society Open Science*, *10*(11), 221328. https://doi.org/10.1098/rsos.221328

Vaidis, D. C., Sleegers, W. W., Van Leeuwen, F., DeMarree, K. G., Sætrevik, B., Ross, R. M., ... & Priolo, D. (2024). A multilab replication of the induced-compliance paradigm of cognitive dissonance. *Advances in Methods and Practices in Psychological Science*, *7*(1). https://doi.org/10.1177/25152459231213375

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences*, *61*, 1340-1341. https://doi.org/10.1002/jps.2600610845