Clinical Psychology: Science and Practice


ISSN: 0969-5893

2021, Vol. 28, No. 2, 226–228
https://doi.org/10.1037/cps0000028

COMMENTARY

# Targeting the Replication Crisis and Improving the Credibility of Research Findings in Clinical Psychology. A Commentary on Pittelkow et al.

Michael P. Hengartner
Department of Applied Psychology, Zurich University of Applied Sciences (ZHAW)



The failure to replicate research findings in psychology is a major scientific issue with important practical implications (Pashler & Harris, 2012), especially with respect to psychological interventions in people with serious mental health problems (Hengartner, 2018). The study by Pittelkow et al. (2021) in this issue is thus timely and commendable. They offer a standardized approach based on Bayes factors (BFs) and subsequent qualitative evaluation to determine which intervention studies in clinical psychology have uncertain or questionable evidential support and are thus in need of replication. Overall, for 42.6% of the statistically significant effects studied, the BFs indicated weak or no evidence for the claimed effect. As I detail below, this is likely an underestimate of the weakness of the evidence, since BFs do not take into account systematic research biases that inflate effect estimates. Nevertheless, this alarming finding corresponds reasonably well with the poor replicability of many research findings as demonstrated in previous studies. In the following, I will discuss some of these studies and offer suggestions as to how we could further improve the credibility of research findings.

## Poor Replicability of Psychological Research Findings

Of 97 statistically significant effects published in three top-tier psychology journals, only 35 (36.1%) were statistically significant in direct replication studies despite increased statistical power to detect these effects (Aarts et al., 2015). Moreover, the average effect size in the replication studies was only half the effect size of the original studies ($r = 0.20$ vs. $r = 0.40$). Likewise, of 83 highly cited articles claiming an effective psychiatric intervention, only 43 (51.8%) were subsequently subjected to an attempt at replication (Tajika et al., 2015). In these 43 replication studies, 16 (37.2%) of the original findings were contradicted, 11 (25.6%) were found to have substantially smaller effect sizes, and only 16 (37.2%) were consistently replicated. As in the study above, the mean effect size in the replication studies was much smaller than in the original studies ($d = 0.31$ vs. $d = 0.72$). The finding that replication studies yield substantially smaller mean effect sizes than the original studies, often close to zero, was confirmed in various other studies (e.g., $d = 0.15$ vs. $d = 0.60$ in the multilab study by Klein et al., 2018). These studies indicate that research findings comprise predominantly insubstantial/trivial effects that were greatly overestimated, or, as Ioannidis put it, that "most published research findings are false" (Ioannidis, 2005).

Until recently, many researchers in psychology did not acknowledge the poor credibility of the scientific evidence, ignored (or downplayed) the uncertainty of many published research findings, and objected to direct replication studies (Ioannidis, 2012; Pashler & Harris, 2012). In fact, independent researchers who conducted direct replication studies routinely faced many obstacles, including a lack of academic recognition/reward, journal editors being uninterested in their replication studies, and encountering resentful researchers/reviewers with vested interests (e.g., the authors of the original studies) who feared damage to their reputation and/or their research field (Ioannidis, 2012; Yong, 2012). Fortunately, major progress has been made in psychology in recent years. Preregistration, open science, and direct replication studies are now increasingly accepted and endorsed by journal editors and researchers alike. Building on the approach proposed by Pittelkow et al. (2021), I want to offer some suggestions as to how we can further improve the detection of questionable research findings in order to determine which intervention studies in clinical psychology are most in need of replication.

## Replications Are Key but Not a Panacea

Poor replicability of research findings is undeniably a major issue undermining the credibility of the scientific literature. Replications are necessary for science to be self-correcting, but even successful replications cannot guarantee improved credibility for published research findings. Both the original findings and the replications could be wrong, a phenomenon Ioannidis termed "perpetuated fallacy" (Ioannidis, 2012). Systematic research biases (e.g., *p*-hacking) can easily lead to repeated successful replications of false positive findings (Ioannidis, 2005; Simmons et al., 2011). This is particularly true for conceptual replications (Ioannidis, 2012; Pashler & Harris, 2012), but it can also occur in direct replications when the methods applied are systematically biased and results are selectively reported (Cuijpers & Cristea, 2016; Hengartner, 2018; Leichsenring et al., 2017). It follows that entire research fields in psychology are possibly built on perpetuated

Michael P. Hengartner https://orcid.org/0000-0002-2956-2969
Correspondence concerning this article should be addressed to Michael P. Hengartner, Department of Applied Psychology, Zurich University of Applied Sciences (ZHAW), P.O. Box 707, CH-8037 Zurich, Switzerland. Email: michaelpascal.hengartner@zhaw.ch

fallacies, that is, repeated (conceptual) replications of artifactual, false positive effects (Ioannidis, 2012); for a concrete example, see Harris et al. (2014).

I thus contend that even large BFs do not necessarily indicate that a reported effect is truly substantial and meaningful. More specifically, a growing body of evidence shows that the effects of psychological interventions are substantially overestimated due to systematic biases in clinical trial methodology, including small sample sizes, a lack of intention-to-treat analyses, inadequate randomization, unblinded outcome assessors, and waiting list control groups (Cuijpers et al., 2010). If we add to these serious methodological limitations the influence of publication bias and selective outcome reporting (Hengartner, 2018), it becomes clear that researchers can quite easily produce evidence that their "therapy is effective, even when it's not" (Cuijpers & Cristea, 2016). Consequently, even when research findings are statistically supported by large BFs (indicating a true effect), the effects may still be insubstantial, that is, methodological artifacts. I thus suggest that the evaluation of the methodology applied to generate the results should be another guiding principle independent of the application of BFs. In the next section, I will go into more detail about which methodological features are most important in my opinion.

## Size of Effect Versus Presence of Effect

As touched on above, even when an observed effect is a true positive (i.e., statistically discernible from a null effect), it may still have very little or no practical relevance (Man-Son-Hing et al., 2002). This is perhaps best demonstrated by the ongoing and unresolved debate regarding whether the treatment effects of antidepressants are clinically (or practically) relevant in the average patient with depression (Hengartner & Plöderl, 2021). Likewise, almost all psychological interventions seem to work very well in carefully preselected participants when compared to a waiting list control group. Treatment effects are even more impressive when the clinical trial is conducted by enthusiastic researchers with preconceived expectations of treatment efficacy and strong allegiance to the investigated therapy (Cuijpers & Cristea, 2016; Leichsenring et al., 2017). However, when compared to placebo or treatment as usual, and when only studies with low risk of bias are considered, and selective reporting is statistically controlled for, then the average treatment effects become disappointingly small and increasingly uncertain (Cuijpers et al., 2010; Cuijpers et al., 2019). It is thus questionable how much benefit, if at all, the average real-world patient truly derives from various empirically supported therapies (see also Sakaluk et al., 2019).

But how can we estimate whether an intervention can really make a meaningful difference in real-world clinical practice? First and foremost, it is essential that researchers fundamentally change the way they evaluate and interpret treatment effects. Instead of comparing an observed effects to a null effect (the standard approach), researchers should test observed effects against a minimally important effect (Hengartner & Plöderl, 2021; Man-Son-Hing et al., 2002). If clinical research ought to translate into improved patient care (which of course is its main objective), then it is insufficient to demonstrate that an intervention has some effect; the effect should also produce a relevant benefit that outweighs harms/costs. Imagine an intensive and time-consuming behavioral intervention in people with obesity that reduces body weight by 200 grams. Even if that effect estimate is precise and statistically discernible from a null effect (0 grams reduction), it is still practically irrelevant. Second, we need to strictly adhere to high-quality clinical trial methodology. If an intervention is designed as a first-line treatment (rather than a second-line or adjunct treatment), it is insufficient to demonstrate efficacy in comparison to a waiting list control group; the intervention should be compared to an established standard of care. It is not in patients' best interest to receive a new therapy that is more effective than doing nothing but inferior to an established therapy. Third, interventions should be tested in large and representative samples, and not in small and narrowly defined samples that have little resemblance to the populations that will later receive the intervention (e.g. interventions studied in students with subthreshold symptoms when designated for psychiatric outpatients with serious mental health problems). Fourth, preregistration of the study according to a comprehensive study protocol and full reporting of all prespecified outcomes is prerequisite to avoid selective reporting and publication bias (Hengartner, 2018; Leichsenring et al., 2017).

## Summary and Conclusion

The study by Pittelkow et al. (2021) presents an interesting approach to determine which intervention studies in clinical psychology are most in need of replication. In addition to calculating BFs, I suggest that poor study quality (or high risk of bias) should be another main criterion for selection of studies with uncertain evidential support. Results from clinical trials that are not preregistered and/or that cannot be checked against a prespecified protocol are questionable and in need of replication regardless of the BFs for the reported effects. This also includes exploratory post-hoc analyses in preregistered studies, which are prone to generate false positive (chance) findings (Ioannidis, 2005; Simmons et al., 2011). I further recommend evaluating a reported effect in comparison to a minimally important effect instead of a null effect, because an intervention effect larger than zero does not imply practical (or clinical) relevance (Hengartner & Plöderl, 2021; Man-Son-Hing et al., 2002). Finally, replications of clinical trial results of course only make sense if replication studies adopt stringent methodology and adhere to high-quality research standards; that is, preregistration according to a comprehensive study protocol, use of large and representative samples, proper randomization, blinding of outcome assessors (study investigators), adequate control groups, intent-to-treat analyses, and full reporting of all prespecified outcomes. Directly replicating research findings from low-quality trials with the same poor methodology will neither advance scientific knowledge nor improve clinical practice. In fact, it rather has the opposite effect, for it will corroborate questionable research findings (perpetuated fallacy) and thus lead to the implementation of interventions with uncertain/questionable effectiveness. But when the field routinely adheres to high research standards, then, and only then, will clinical psychology research become self-correcting through replication, ultimately generating reliable, credible, and practically meaningful scientific evidence that truly helps people with mental health problems.

## References


Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., Babel, M., Bahník, Š., Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Boorsboom, D., Bosch, A., Bosco, F. A., . . . Zuni, K. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Cuijpers, P., & Cristea, I. A. (2016). How to prove that your therapy is effective, even when it is not: A guideline. *Epidemiology and Psychiatric Sciences*, *25*(5), 428–435. https://doi.org/10.1017/S2045796015000864

Cuijpers, P., Karyotaki, E., Reijnders, M., & Ebert, D. D. (2019). Was Eysenck right after all? A reassessment of the effects of psychotherapy for adult depression. *Epidemiology and Psychiatric Sciences*, *28*(1), 21–30. https://doi.org/10.1017/S2045796018000057

Cuijpers, P., van Straten, A., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010). The effects of psychotherapy for adult depression are overestimated: A meta-analysis of study quality and effect size. *Psychological Medicine*, *40*(2), 211–223. https://doi.org/10.1017/S0033291709006114

Harris, C. R., Pashler, H., & Mickes, L. (2014). Elastic analysis procedures: An incurable (but preventable) problem in the fertility effect literature. Comment on Gildersleeve, Haselton, and Fales (2014). *Psychological Bulletin*, *140*(5), 1260–1264. https://doi.org/10.1037/a0036478

Hengartner, M. P. (2018). Raising awareness for the replication crisis in clinical psychology by focusing on inconsistencies in psychotherapy research: How much can we rely on published findings from efficacy trials? *Frontiers in Psychology*, *9*, 256. https://doi.org/10.3389/fpsyg.2018.00256

Hengartner, M. P., & Plöderl, M. (2021). Estimates of the minimal important difference to evaluate the clinical significance of antidepressants in the acute treatment of moderate-to-severe depression. *BMJ Evidence-Based Medicine*. Advance online publication. https://doi.org/10.1136/bmjebm-2020-111600

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

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

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Vega, D., Aveyard, M., Axt, J., Babaloia, M., Bahník, Š., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Bocian, K., Brandt, M., Busching, R., Cai, H., . . . Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Leichsenring, F., Abbass, A., Hilsenroth, M. J., Leweke, F., Luyten, P., Keefe, J. R., Midgley, N., Rabung, S., Salzer, S., & Steinert, C. (2017). Biases in research: Risk factors for non-replicability in psychotherapy and pharmacotherapy research. *Psychological Medicine*, *47*(6), 1000–1011. https://doi.org/10.1017/S003329171600324X

Man-Son-Hing, M., Laupacis, A., O'Rourke, K., Molnar, F. J., Mahon, J., Chan, K. B., & Wells, G. (2002). Determination of the clinical importance of study results. *Journal of General Internal Medicine*, *17*(6), 469–476. https://doi.org/10.1046/j.1525-1497.2002.11111.x

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536. https://doi.org/10.1177/1745691612463401

Pittelkow, M. M., Hoekstra, R., Karsten, J., & van Ravenzwaaij, D. (2021). Replication target selection in clinical psychology: A Bayesian and qualitative re-evaluation. *Clinical Psychology: Science and Practice*, *28*(2), 210–221.

Sakaluk, J. K., Williams, A. J., Kilshaw, R. E., & Rhyner, K. T. (2019). Evaluating the evidential value of empirically supported psychological treatments (ESTs): A meta-scientific review. *Journal of Abnormal Psychology*, *128*(6), 500–509. https://doi.org/10.1037/abn0000421

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Tajika, A., Ogawa, Y., Takeshima, N., Hayasaka, Y., & Furukawa, T. A. (2015). Replication and contradiction of highly cited research papers in psychiatry: 10-year follow-up. *The British Journal of Psychiatry*, *207*(4), 357–362. https://doi.org/10.1192/bjp.bp.113.143701

Yong, E. (2012). Replication studies: Bad copy. *Nature*, *485*(7398), 298–300. https://doi.org/10.1038/485298a


Received March 3, 2021
Accepted March 26, 2021 ■