Underpowered Studies and Overrepresented Significant Findings in Educational Psychology:
A Comprehensive Examination of Empirical Evidence

Abstract

Most psychological studies are underpowered, and most published findings are false positives (Ioannidis, 2005). Such studies are prone to replication failures. Meta-analyses that synthesize questionable research findings may yield biased results. The current study aims to thoroughly examine the credibility of meta-analyses of educational psychology research. To assess the replicability and robustness of educational psychology findings, we investigate power and potential false outcomes for selected meta-analyses recently published in leading journals. The preliminary results suggest that meta-analyses of experimental studies published in the years 2020 and 2021 are severely underpowered. We wholeheartedly encourage educational psychologists to more actively appreciate the importance of power analysis in psychological research and reconsider their approach to determining sample sizes.

Keywords: median retrospective power, false positive rate, replication crisis, meta-analysis

Underpowered Studies and Overrepresented Significant Findings in Educational Psychology: A Comprehensive Examination of Empirical Evidence

Objectives

The current study examines the credibility of meta-analyses in educational psychology from two aspects: statistical power and true-positive rates. The first objective is to calculate the median retrospective power of meta-analyses recently published in top-tier educational psychology journals. The second objective involves examining whether the calculated power of educational psychology studies changed substantially over the five-year period from 2017 to 2022. The third objective is to estimate the true-positive probability of meta-analyses in educational psychology using a Bayesian approach. We will explore potential factors that lead to low power and provide recommendations for improving replication success in educational psychology.

Literature and theoretical framework

Low statistical power and excess false positive findings

Over the past decade, the field of psychology has gone to great lengths to scrutinize well-established findings via large-scale, preregistered multi-lab replication projects. The results revealed a significant replication crisis — only about 36% of replicated studies yielded statistically significant results and replicated effect sizes were approximately half as large as initial findings indicated (Open Science Collaboration, 2015).

An observed treatment effect can only be considered useful and reliable for research and clinical purposes if the effect size is not overestimated and represents a true positive. Statistical power plays a key role in ensuring the reliability, precision, and replicability of an effect size (Fraley & Vazire, 2014). In psychology, the average power is estimated to be 35-36% and only 8% of studies are estimated to be adequately powered (Bakker et al., 2012; Stanley et al., 2018). Such studies produce unreliable findings prone to replication failures. One serious issue associated with low power is increased false positive rates (Button et al.,

2013; Pashler & Harris, 2012). Theoretically, the proportion of positive findings in the literature should be roughly equal to the average power (Nuijten et al., 2015). This means that apart from a small number of false positive findings (i.e., the Type I error), only about 35% of psychology studies can discover statistically significant effects — on those occasions when they're *actually* true, of course. However, a large survey on published studies conducted by Fanelli (2010) found that a staggering 95.1% of Psychology/Psychiatry studies reported significant results. When studies are alarmingly underpowered, a substantial proportion of findings in psychological literature are likely to be false. Moreover, they're bound to have inflated effect size estimates (Fraley & Vazire, 2014).

False positive meta-analyses due to low power

Meta-analysis is considered the gold standard for quantitative synthesis of research findings. However, when a field is dominated by underpowered studies, meta-analytic findings can be strongly biased. When the primary studies included in a meta-analysis are underpowered, meta-analytic studies also face the risk of high false positive rates (Pereira & Ioannidis, 2011; Stanley et al., 2022a) and inflated effect size estimates (Brand et al., 2008; Friese & Frankenbach, 2020).

Median retrospective power (MRP)

Median retrospective power (MRP) indicates whether a typical study in a meta-analysis is sufficiently powered to discover a true positive effect and if its direct replication can be successful (Stanley et al., 2022b). Using simulation studies and empirical data from Many Labs 2 replication project (Klein et al., 2014), Stanley et al. (2022a) demonstrated that MRP can serve as a valuable indicator of meta-analysis credibility. When MRP exceeds 50% or 60%, the probability that a meta-analysis result is a false positive decreases considerably. In such cases, meta-analysis findings are far more likely to be trustworthy. The formula to calculate MRP is

$$MRP = 1 - N(1.96 - \frac{|UWLS|}{SE_m}), \tag{1}$$

where $N()$ is the cumulative normal probability, SEm is the median of the standard errors of study effects, and UWLS is the unrestricted weighed least squares — a meta-analytic effect size estimate. UWLS is capable of utilizing information of the entire research area while effectively correcting for selective reporting bias (Stanley & Doucouliagos, 2015, 2017). Given the strong probability that a meta-analysis may simultaneously feature substantial heterogeneity and publication bias, Stanley and Doucouliagos (2015) argued that the UWLS can deliver more accurate estimates than those provided by the fixed-effect or random-effects models in conventional meta-analysis.

True positive rate (credibility) of meta-analysis

Ioannidis (2008) used a Bayesian approach to calculate the proportion of true positives in meta-analyses, thereby evaluating their credibility. To estimate the credibility of a meta-analysis, we introduce an equation that describes the relationship between the Bayes factor and the posterior odds:

$$Posterior\,odds = Bayes\,factor \times Prior\,odds. \tag{2}$$

The prior odds represents the initial beliefs assigned to different hypotheses before any new evidence is considered. The posterior odds represent the updated beliefs or probabilities after new evidence is incorporated into the analysis. The Bayes factor, in essence, is the probability ratio of data given hypotheses (Rouder & Morey, 2011). More specifically, it compares the relative probability for two competing models by examining the marginal likelihoods of each model (i.e., the probability of the observed data given the corresponding model).

In their studies, Ioannidis (2005) and Pereira and Ioannidis (2011) denoted C as the posterior probability of an true effect (with $\frac{C}{1-C}$ representing the posterior odds), R as the prior odds, and B as Bayes factor. The formula for credibility is therefore defined as

$$C = \frac{R}{R + B}. \tag{3}$$

When we calculate C for a meta-analysis, we need to assign a possible value to R,

depending on the degree of confidence in the hypothesis being tested. There are numerous ways to estimate B. The authors estimated B based on a method proposed by Spiegelhalter et al. (2004, p.131). For each selected meta-analysis, B is estimated as

$$B = \sqrt{1 + \frac{\pi\theta_A}{\sigma^2(\hat{\theta})}} \times e \left[\frac{\frac{-\hat{\theta}}{\sigma(\hat{\theta})}}{2(1 + \frac{\sigma^2(\hat{\theta})}{\pi\theta_A^2})} \right], \tag{4}$$

where $\hat{\theta}$ and $\sigma^2(\hat{\theta})$ are the summary effect size and its variance, respectively. θ_A is the mean effect size under the alternative model, if the effect truly exists. When calculating B, we need to specify θ_A with a predetermined value that is consistent with empirical evidence.

Methods

Data collection

Meta-analyses provides the best source of evidence for retrospective power analysis. We systematically searched, collected, and screened eligible meta-analyses published in leading peer-reviewed educational psychology journals. We chose journals based on two journal rankings: Google Scholar and Scimago Journal & Country Rank. The included journals were Educational Psychologist, Educational Psychology Review, Journal of Educational Psychology, Contemporary Educational Psychology, and British Journal of Educational Psychology. We searched meta-analyses in two databases: Education Source and ERIC. We used the following search terms: “meta-analysis (Title) AND educational psycholog* (All Fields) AND (Year of Publication)”. We established and adhered to the following inclusion criteria:

1. The study must be a conventional meta-analysis (not a network meta-analysis).
2. The meta-analysis must provide sufficient data to calculate MRP and Bayes factor
3. The meta-analysis must be written in English.
4. The focus of the meta-analysis must be related to educational psychology.

Data extraction and analysis

We extracted two types of effect sizes: standardized mean difference (SMD; Cohen's d and Hedges' g) and correlation coefficient (r). When calculating MRP using SMDs, if a selected meta-analysis provided raw data from primary studies, we extracted the sample size, group mean, and standard deviation for both control and treatment groups in each primary study within the meta-analysis. If raw data were not provided, we extracted the effect sizes and their associated standard errors (SE). Unexpectedly, many studies neither reported raw data nor SEs. In these instances, we derived SEs from the 95% CIs of the effect sizes. To calculate MRP using correlation coefficients, we extracted r, sample size, and SE in each primary study within the meta-analysis.

For the calculation of the Bayes factor, we will extract the summary effect sizes and their variances from all meta-analyses. We will establish a priori θ_A values that corresponds to a mean SMD of 0.3 and a mean correlation of 0.2, respectively. These values are roughly consistent with a large survey of 200 meta-analyses conducted by Stanley et al. (2018). We consider two different values for the prior odds, R, as suggested by Pereira and Ioannidis (2011): 0.5 and 0.1. When R is 0.5, the prior probability of a tested effect being true is 33% (1:2 odds). When R is 0.1, the prior probability of a tested effect being true is 9% (1:10 odds).

Software

All statistical programming and analyses are conducted in the R programming software (R Core Team, 2022).

Preliminary results

We collected 18 meta-analyses published in 2021 and 21 meta-analyses published in 2020. We were able to calculate MRP in 7 out of 18 studies published in 2021, which contributed 19 summary effect sizes. We were able to calculate 7 out of 21 studies published in 2020, which contributed 9 summary effect sizes.

We found that the median SMD estimated by the UWLS method was 0.33, which

was considered a small effect size. The median correlation estimated by the UWLS method was 0.32, which was considered a medium effect size.

The median MRP of the studies included in SMD-based meta-analyses was estimated to be 31%, indicating that most of the experiment-based studies were inadequately powered. A median power of 31% means that the typical replication study that uses the typical care in duplicating research protocols, conditions, and methods and typical sample sizes will have only a 31% chance of finding a statistically significant effect in the expected direction.

The median MRP of the studies included in correlation-based meta-analyses is estimated to be 95%, indicating that most of the correlation-based studies were highly powered.

The remaining data analyses are expected to be completed during Fall 2023.

Scientific significance

Despite repeated calls to improve statistical power, the persistence of underpowered studies and high false positive rates remains a major issue in social science research (Bakker et al., 2012). The extent to which these issues affect educational psychology has not received sufficient attention they deserve. To the best of our knowledge, this study is the first to comprehensively investigate such issues in the field. Its completion will help the academic community raise awareness of the systemically underpowered research in top-tier journals, and also increase cognizance of the degree to which existing findings previously deemed "significant" are more likely spurious or biased in favor of inflated effects. Research demonstrates that a surprisingly large number of psychologists have flawed intuitions about power and fail to conduct power analyses appropriately in their research (Bakker et al., 2016). As such, we urge educational psychologists to more actively appreciate the gravity of power analysis in psychological research, in addition to reevaluating their sample size planning criteria. Doing so will ensure sufficient power to identify true, replicable effect sizes.

References

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' intuitions about power in psychological research. *Psychological science*, *27*(8), 1069–1077.

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554.

Brand, A., Bradley, M. T., Best, L. A., & Stoica, G. (2008). Accuracy of effect size estimates from published psychological research. *Perceptual and motor skills*, *106*(2), 645–649.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, *14*(5), 365–376.

Fanelli, D. (2010). "positive" results increase down the hierarchy of the sciences. *PloS one*, *5*(4), e10068.

Fraley, R. C., & Vazire, S. (2014). The n-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS one*, *9*(10), e109019.

Friese, M., & Frankenbach, J. (2020). P-hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*, *25*(4), 456.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Ioannidis, J. P. (2008). Effect of formal statistical significance on the credibility of observational associations. *American journal of epidemiology*, *168*(4), 374–383.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahnık, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., et al. (2014). Investigating variation in replicability. *Social psychology*.

Nuijten, M. B., Van Assen, M. A., Veldkamp, C. L., & Wicherts, J. M. (2015). The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Review of General Psychology*, *19*(2), 172–182.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536.

Pereira, T. V., & Ioannidis, J. P. (2011). Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *Journal of clinical epidemiology*, *64*(10), 1060–1069.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org

Rouder, J. N., & Morey, R. D. (2011). A bayes factor meta-analysis of bem's esp claim. *Psychonomic Bulletin & Review*, *18*, 682–689.

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation* (Vol. 13). John Wiley & Sons.

Stanley, Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological bulletin*, *144*(12), 1325.

Stanley & Doucouliagos, H. (2015). Neither fixed nor random: Weighted least squares meta-analysis. *Statistics in medicine*, *34*(13), 2116–2127.

Stanley & Doucouliagos, H. (2017). Neither fixed nor random: Weighted least squares meta-regression. *Research synthesis methods*, *8*(1), 19–42.

Stanley, Doucouliagos, H., & Ioannidis, J. P. (2022a). Beyond random effects: When small-study findings are more heterogeneous. *Advances in Methods and Practices in Psychological Science*, *5*(4), 1–11.

Stanley, Doucouliagos, H., & Ioannidis, J. P. (2022b). Retrospective median power, false positive meta-analysis and large-scale replication. *Research Synthesis Methods*, *13*(1), 88–108.