

Estimating the replicability of quantitative engineering education research

J. Buckley ^{a,b,1}, S. O'Connor ^c, J. Hallström ^d

^a Technological University of the Shannon: Midlands Midwest, Athlone, Ireland

^b Linköping University, Norrköping, Sweden, 0000-0002-8292-5642

^c University of Limerick, Limerick, Ireland, 0000-0001-5069-5953

^d Linköping University, Norrköping, Sweden, 0000-0003-0829-3349

Conference Key Areas: *Strengthening EER (Engineering Education Research) as a research field. Dialogue between engineering and society – effects on education.*

Keywords: *Engineering education research, Replicability, Z-curve analysis, Publication bias, Research credibility.*

Abstract

This paper provides the first systematic estimate of replicability in engineering education research. Estimating and improving replicability is essential for the provision of strong evidence to support improvements in educational practice. Using z-curve 3.0 to analyse 1,366 statistical tests extracted from 5,551 full-text articles published in academic journals between 2016 and 2026, the study assesses the field's credibility in light of the "replication crisis". Findings indicate an estimated expected replication rate of 71% for exact replication attempts of significant observations, and an actual replication prediction of 56% considering difficulties in conducting exact replication studies. While these results mirror findings in technology education, the analysis reveals significant evidence of publication bias and *p*-hacking across the full dataset. Notably, the *Journal of Engineering Education* showed evidence of selection bias rather than *p*-hacking, though this is potentially due to its reputation as a "premier" venue. To enhance quantitative engineering education research credibility, it is recommend aiming to increase sample sizes to achieve 80% statistical power and adopt pre-registration practices to mitigate reporting bias.

¹ Corresponding Author
J. Buckley
Jeffrey.buckley@tus.ie

1 Introduction

Discussions pertaining to the enhancement of engineering education research have been ongoing for over a decade, but have focused more on the use of theory (Beddoes, 2022), qualitative (Bernhard & Baillie, 2016) and mixed methods research (Agarwal & Raju, 2025), and the conduction of systematic reviews and use of general open science principles (Power, 2021b, 2021a) than on quantitative research. It seems pertinent to expand these discussions to quantitative research given its frequency in engineering education research (Loui & Borrego, 2019; Satya Putra et al., 2023) and in light of calls for more advanced quantitative research in the field to enable engagement with more complex and nuanced research questions (Borrego et al., 2009; Satya Putra et al., 2023).

Traditionally, there has been a focus on ‘rigor’ in empirical engineering education research (Borrego, 2007), with the Rigorous Research in Engineering Education project adopting the definition of rigor from the US National Research Council (Committee on Scientific Principles for Education Research, 2002). This suggested that rigorous engineering education research should pose significant questions that can be answered empirically, link research to relevant theory, use methods that permit direct investigation of the question, provide a coherent and explicit chain of reasoning, replicate and generalize across studies, and disclose research to encourage professional scrutiny and critique.

Following these calls for increased rigor in engineering education research, the social sciences more broadly became aware of what has since been termed the “replication crisis” (Shrout & Rodgers, 2018). From the early 2010s, initially in social psychology (Pashler & Wagenmakers, 2012) though extending to other fields (cf. Baker, 2016), the replication crisis was characterised by replication failures, awareness of the impact of questionable research practices on published results and their prevalence, as well as cases of academic fraud. This replication crisis sparked a credibility revolution (Vazire, 2018), where credibility is a broader concept than rigour relating to the general trustworthiness of research (Nosek et al., 2026), and in light of this, at least for quantitative work, engineering education research should perhaps adopt this broader aspiration of credibility than rigor.

One aspect of credibility, coincidentally overlapping with rigor, is replicability. Replicability is related to observing consistent findings pertaining to the same research question in independent studies and has significant implications to the translation of observations to practice. Where there is a goal of generalisability, successful replication attempts of original studies add to the validity of inferences and provide stronger empirical support for translating educational insights into practice, such as to pedagogy and policy. The need for replication research in education has been recently identified and emphasised (Banerjee, 2022; Cai, 2022; Morrison, 2022; Perry et al., 2022; Perry & See, 2022; Wiliam, 2022), and though the frequency of replication studies is growing it remains very small. Only 0.2% of published education research studies between 2011 and 2020 (Perry et al., 2022) and only 0.16% of published technology education studies between 1989 and 2026 (Buckley, 2026) were intentional replication studies. Replicability can be assessed for individual studies through replication attempts, and at field level. At field level, one approach is to conduct several replication studies, and use the success rate as an estimate of field-wide replicability (Tyner et al., 2026). This is resource intensive,

however alternative meta-analytic methods can statistically estimate replicability. In the cognate area of technology education, Buckley et al. (2023) performed a z-curve analysis (Bartoš & Schimmack, 2022; Brunner & Schimmack, 2020), estimating replicability of technology education research as 81% for exact replication attempts of observed significant results, and 56% for actual replication studies given the challenges in exact replication. There has been no analysis of the replicability of quantitative engineering education research to date, and such estimates are needed to guide on improving this dimension of research credibility, e.g., do sample sizes need to increase? This paper reports a z-curve analysis of quantitative engineering education research published in academic journals over the last 10 years to address this need.

2 Methodology

2.1 Approach

To estimate the replicability of quantitative engineering education research, this study involved the use of z-curve analysis (Bartoš & Schimmack, 2022; Brunner & Schimmack, 2020), specifically z-curve 3.0 (Schimmack, 2025a) implemented through RStudio with R version 4.2.2 (The R Foundation for Statistical Computing, 2022). Z-curve is a meta-analytic tool designed to examine the credibility of a body of literature specifically with respect to replicability and statistical power, which is the probability of detecting an effect in a statistical test performed on a sample under the assumption that there is an effect to be detected in reality. Replicability is estimated through z-curve analyses by modelling the distribution of z-scores, which are converted from p-values in reported hypothesis tests (e.g., $p = .05$ equates to $z \approx 1.96$ [two-tailed]). From this, observed power can be estimated, and estimates can be derived pertaining to the likely amount of missing/unpublished studies, the amount of bias, and likely replication rates. A utility of z-curve is that it can estimate replicability in heterogeneous datasets (Bartoš & Schimmack, 2022), i.e., studies which use different quantitative methodologies, or are responding to different research questions. Typically, z-curve is applied to studies with some common feature. In this study, that common feature is publication in engineering education research journals within the last decade (2016 to March 2026).

Outputs from z-curve 3.0 include the Observed Discovery Rate (ODR), which is the proportion of significant results ($p < .05$) observed in a dataset, the Expected Discovery Rate (EDR), which is an estimate of the proportion of significant results that would be observed if there were no selection bias in published studies, the Expected Replication Rate (ERR), which is an estimate of the proportion of studies with significant results that would replicate in an exact replication study, and the False Discovery Rate (FDR), which is an estimate of the proportion of studies which are false positive results. Furthermore, given exact replication is typically not feasible, Schimmack (2022) suggests that by taking the average of the EDR and ERR values, an Actual Replication Prediction (ARP) rate can be computed as an estimate of the likely success rate in actual replication studies. Finally, z-curve 3.0 includes a test of Excessive Just Significance (EJS), which examines the distribution of reported z-values and checks whether the resulting model can account for a “just significant” region (in this study, set at $z = 1.96$ to 2.6 , i.e., $p \approx .05$ to $.01$) of reported statistical tests. A significant EJS test result suggests publication bias, and this same test can then be applied to the distribution of significant z-scores only to check if that

model can account for z-scores in determined just significant regions (in this study, both $z = 1.96$ to 2.6 , i.e., $p \approx .05$ to $.01$ and $z = 1.96$ to 2.4 , i.e., $p \approx .05$ to $.015$ were selected), with results for these tests used to infer if identified bias relates to either p -hacking or selection bias (Schimmack, 2025b).

2.2 Dataset

To identify articles published in engineering education research journals, initially a Scopus search was performed on 26th February 2026. The intention of this search was to identify sources (academic journals) relating to engineering education. The specific search string was:

SRCTITLE ("engineering education") AND PUBYEAR > 2015

This search however returned works published in other source types, such as conferences and books, which were excluded from this analysis to minimise the likelihood of including duplicate statistical results which could occur for example from a work-in-progress conference paper subsequently being published in an academic journal. Therefore, in Scopus, limitations were placed on the search relating to document type (limited to "article"), publication stage (limited to "final"), and source type (limited to "journal"). The final Scopus search string for reproducibility, which produced 6,498 results on 26th February 2026 was:

SRCTITLE ("engineering education") AND PUBYEAR > 2015 AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (PUBSTAGE, "final")) AND (LIMIT-TO (SRCTYPE, "j"))

The next step required the collation of full texts for identified articles. This process was undertaken between 26th February 2026 and 9th March 2026, utilising university library subscriptions and databases such as Google Scholar and ResearchGate. While searching for full texts, it became apparent that the Scopus search had not identified all issues within each journal, and the journal *SEFI Journal of Engineering Education Advancement* had not been identified. Therefore, each Scopus identified journal and the *SEFI Journal of Engineering Education Advancement* were manually searched. The Zotero reference management software, inclusive of its PDF resolver and Chrome browser plugin, was used to support full text download. Publications other than articles, such as editorials, may have been added to the dataset, however this is not problematic as only quantitative statistical results are extracted for analysis, which typically would not be present in such publications. The final search included a total of 7,099 articles. Of the total of 7,099 articles, 5,551 full texts were found (78.2%) and 1548 were not found (21.8%). Most not found texts ($n = 1,249$) came from the *International Journal of Engineering Education* as the authors did not have a journal subscription, so the final dataset should be considered as not reflective of the work published in this journal (Table 1).

Table 1. Study dataset comprising of $n = 5,551$ full texts (78.2%) of 7,099 articles across 16 engineering education journals.

Journal	Found PDF	Not Found PDF	Total Articles	Full Texts (%)
AEE	257	0	257	100%
AJEE	112	0	112	100%

CEE	247	21	268	92%
CAEE	1218	0	1218	100%
EJEE	647	0	647	100%
GJEE	306	0	306	100%
IJCEELLL	153	209	362	42%
IJEE	12	1249	1261	1%
IJMEE	279	0	279	100%
JCEE	75	65	140	54%
JEE	361	0	361	100%
JEET	1487	4	1491	100%
JPCEER	122	0	122	100%
JPIEEP	177	0	177	100%
SEE	63	0	63	100%
SJEEA	23	0	23	100%

Note. AEE = Advances in Engineering Education. AJEE = Australasian Journal of Engineering Education. CEE = Chemical Engineering Education. CAEE = Computer Applications in Engineering Education. EJEE = European Journal of Engineering Education. GJEE = Global Journal of Engineering Education. IJCEELLL = International Journal of Continuing Engineering Education and Life-Long Learning. IJEE = International Journal of Engineering Education. IJMEE = International Journal of Mechanical Engineering Education. JCEE = Journal of Civil Engineering Education (formerly JPIEEP). JEE = Journal of Engineering Education. JEET = Journal of Engineering Education Transformations. JPCEER = Journal of Pre-College Engineering Education Research. JPIEEP = Journal of Professional Issues in Engineering Education and Practice (became JCEE). SEE = Studies in Engineering Education. SJEEA = SEFI Journal of Engineering Education Advancement.

The final step in preparing data for analysis required the extraction of statistical test information from the 5,551 full texts. There are two main approaches to doing this once full texts are identified. Focal coding would involve full text review of each article, and manually extracting statistical test information from focal hypothesis tests (e.g., Bartoš, 2024). For this study, automated extraction of statistical test information from the full texts was used, similar to Buckley et al. (2023) Gupta and Bosco (2023), Soto and Schimmack (2025), and Schimmack and Bartoš (2023), specifically using the statcheck R package (Epskamp & Nuijten, 2018). This resulted in 1,366 extracted statistical tests across 13 of the 16 journals (Table 2).

Table 2. Descriptive statistics for quantities of statistical tests extracted per journal using the statcheck R package (Epskamp & Nuijten, 2018).

Journal	Unique articles	Total extracted tests	Mean tests per article	Median tests per article	Min tests per article	Max tests per article
AEE	18	72	4	3	1	13
AJEE	10	46	4.6	3	1	18
CAEE	56	388	6.93	4	1	49
CEE	10	32	3.2	3	1	6
EJEE	57	376	6.6	4	1	35
GJEE	3	11	3.67	4	2	5
IJCEELLL	3	18	6	5	1	12
IJEE	1	3	3	3	3	3
IJMEE	4	27	6.75	2.5	1	21
JCEE	1	9	9	9	9	9
JEE	42	280	6.67	3	1	59

JEET	29	102	3.52	3	1	13
JPCEER	2	2	1	1	1	1

3 Results

Z-curve analyses were conducted on all data ($n_{\text{tests}} = 1,366$) as well as at journal level ($n_{\text{journals}} = 9$). Due to insufficient data, z-curves could not be estimated for the GJEE, IJEE, JCEE, and JPCEER journals, though the extracted data from these sources was still included in the full dataset. Following the procedure described by Schimmack (2025b), four z-curve analyses were conducted in each instance with 500 bootstraps to estimate 95% confidence intervals. The first z-curve was used to test for bias with an EJS test examining the distribution of all z-scores and testing whether the model could explain the frequency of z-scores between $z = 1.96$ and $z = 2.6$. A significant result ($p < .05$) would indicate bias. The second and third z-curves were performed to test for p -hacking with an EJS test, by examining the distribution of all significant z-values and testing whether the models could explain the frequency of z-scores between $z = 1.96$ and $z = 2.6$, and between $z = 1.96$ and $z = 2.4$ respectively. As two tests were conducted, the alpha level for statistical significance is adjusted, such that a significant result ($p < .025$) would indicate evidence of p -hacking. A final model was performed to estimate EDR, ERR, FDR, and ARP rates, by modelling the distribution of all z-scores and examining how this relates to all significant z-scores ($z > 1.96$). Full results are presented in **Chyba! Nenalezen zdroj odkazů.**, with an illustrative z-curve, the final model for all data, presented in Figure 1.

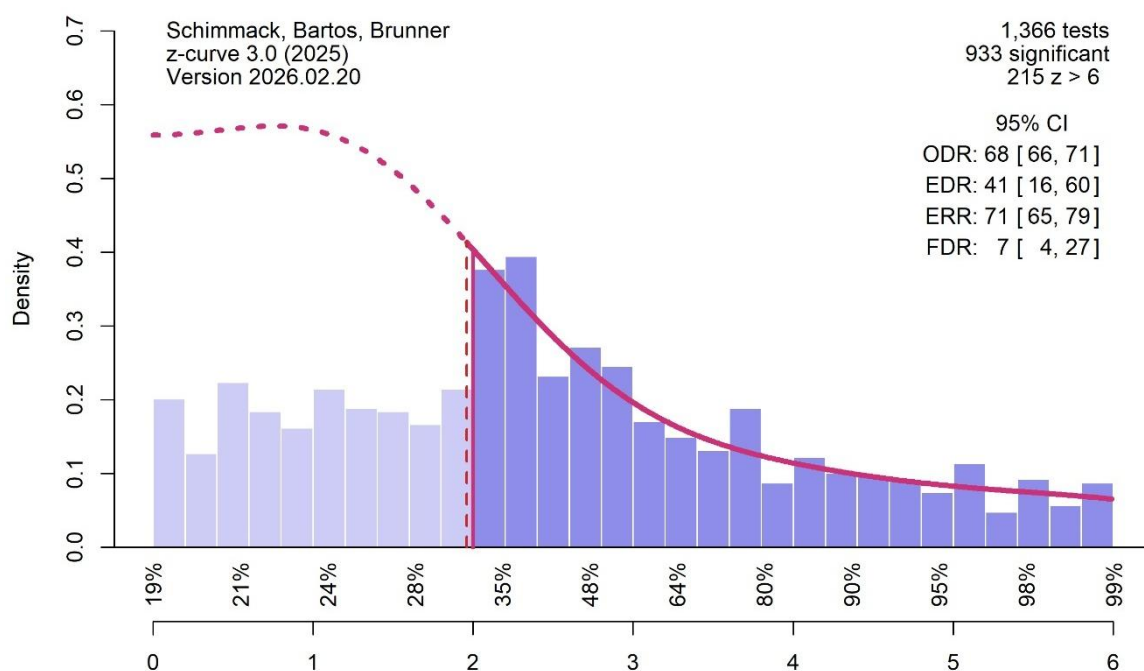


Figure 1. Final z-curve for all data.

Figure 1 shows a histogram of all z-scores, with the solid red line representing the z-curve fit to the distribution of observed significant z-scores and the dashed red line showing the expected distribution without selection bias. The model overestimates the number of non-significant results, indicating bias in the literature. Percentage

values on the x-axis indicate local statistical power estimates at various levels of significance.

4 Discussion and conclusions

The results indicate that for all data, there is an estimated success rate for exact replication attempts of significant observations of 71%, 95% CI [65%, 79%], though this should be considered in light of the difficulties in performing exact replication and thus the actual replication prediction rate of 56% may be a more appropriate estimate. These are very similar estimates to a previous z-curve analysis by Buckley et al. (2023) in technology education research, who for published works between 1983 and 2020 observed an ERR of 81% and an ARP of 56%. At individual journal level, there is a wide ARP range of 17% to 91%, but considering the CAEE, EJEE, and JEE as the journals with most data, the ARP values range from 43% to 65%.

In terms of bias testing, there is evidence of publication bias in the full dataset, and this can at least partially be explained by the use of *p*-hacking, which is defined as “a compound of strategies targeted at rendering non-significant hypothesis testing results significant” (Stefan & Schönbrodt, 2023, p. 1), often done to improve publishability of work. The implication of *p*-hacking is the inflation of false positive results and, consequently, the lower replicability of studies (Abdol & Wicherts, 2021), though not all researchers engaging in the practice may understand or appreciate this (Stefan & Schönbrodt, 2023). At journal level, only the JEE shows evidence for bias, but not *p*-hacking, suggesting that the observed bias is selection or reporting bias, i.e., the preferential publication of statistically significant results over null results. This can skew the literature base in a way suggestive that interventions are more likely to work or variables are more likely to be related than they are, as readers are not exposed to the instances where no evidence of effect is observed. However, what cannot be determined is whether the bias takes place at journal level (the journal favouring significant results), author level (authors choosing the journal for publishing significant findings over null findings) or both. In this case, the JEE does hold a reputation as being the “premier” journal in the field (e.g., Pitterson et al., 2020) and thus there is a likelihood that on observing a significant result, researchers may look to the JEE in the first instance. No other journal shows evidence for bias.

The implications of this analysis are that, even though it is reasonably good, statistical power could be improved through increased sample sizes. As shown in Figure 1, non-significant results ($z < 1.96$) are estimated to have low power ($\approx 19\%$ - 30%), and the recommendation for 80% power (cf. Cohen, 1988) is not achieved until $z \approx 4$. Further, at least with respect to the JEE, though the implication should be considered field-wide, there does appear merit in increasing the practice of pre-registration (cf. Nosek et al., 2018), at least for quantitative work, to remove and prevent potential selection bias and work towards a literature base impartial to results and instead based on methodological credibility. Future work should also consider robustness analyses (Aczel et al., 2026) in the first instance to support generalised claims and reproducibility checks (Miske et al., 2026) to further aid credibility. These implications should be considered in light of the study limitations. The estimated replicability rates are general rather than identifying specific studies. There are other engineering education research journals not indexed by Scopus, non-English dissemination, grey literature, and empirical works published in other sources such as STEM education journals. Additionally, full texts were not found for

all articles, notably for IJEE, and data extraction was automated by statcheck which requires APA formatting that may not be adhered to across all engineering education journals. As such, there is potential sampling bias limiting generalisability.

Table 3. Z-curve results. 95% confidence intervals estimated through 500 bootstraps. A Kernel Density model was fit to the z-curves computed to perform estimated just significant (EJS) tests. An Expectancy Maximisation model was fit to the z-curves computed to estimate Observed Discovery Rate (ODR), Expected Discovery Rate (EDR), Expected Replication Rate (ERR), Actual Replication Prediction (ARP) and False Discovery Risk (FDR).

Journal	Total extracted tests	EJS (z = 1.96 - 2.6)	EJS (z > 2.6)	EJS (z > 2.4)	ODR (95% CI)	EDR (95% CI)	ERR (95% CI)	ARP	FDR (95% CI)
All data	1366	.001*	1.00	.007**	.683 (.658, .708)	.414 (.162, .596)	.712 (.646, .787)	.563	.074 (.036, .271)
AEE	72	.166	.360	.103	.625 (.513, .737)	.150 (.050, .569)	.552 (.333, .757)	.351	.298 (.040, 1.00)
AJEE	46	.504	.044*	.049*	.500 (.356, .644)	.169 (.050, .863)	.649 (.380, .891)	.409	.258 (.008, 1.00)
CAEE	388	.172	.640	.364	.634 (.586, .682)	.502 (.187, .822)	.793 (.695, .892)	.647	.052 (.011, .229)
CEE	32	.833	.991	.984	.625 (.457, .793)	.838 (.192, 1.00)	.875 (.655, .989)	.856	.010 (.000, .221)
EJEE	376	.193	1.00	.078	.718 (.673, .764)	.312 (.115, .657)	.720 (.624, .826)	.516	.116 (.028, .404)
IJCEELL	18	.285	.970	.230	.667 (.449, .884)	.116 (.050, .584)	.226 (.025, .606)	.171	.401 (.037, 1.00)
IJMEE	27	.132	.981	.997	1.00 (1.00, 1.00)	.892 (.495, 1.00)	.933 (.778, 1.00)	.912	.006 (.000, .054)
JEE	280	.002*	1.00	.780	.686 (.631, .740)	.288 (.066, .468)	.578 (.474, .696)	.433	.130 (.060, .748)
JEET	102	.416	.710	.607	.794 (.716, .873)	.852 (.389, .987)	.892 (.770, .973)	.872	.009 (.001, .083)

Note. * $p < .05$. ** $p < .025$. EJS = Excessive Just Significance test. ODR = Observed Discovery Rate. EDR = Expected Discovery Rate. ARP = Actual Replication Prediction. FDR = False Discovery Risk. CI = Confidence Interval. An α value of .05 is used as a threshold for inferring statistical significance in the general test of bias, EJS ($z = 1.96 - 2.6$). To account for multiple comparisons, an adjusted α value of .025 is used as a threshold for inferring statistical significance both tests of p -hacking, EJS ($z > 2.6$) and EJS ($z > 2.4$).

5 References

- Abdol, A. M., & Wicherts, J. M. (2021). *Science abstract model simulation framework* (Zy29t_v1). PsyArXiv. <https://doi.org/10.31234/osf.io/zy29t>
- Aczel, B., Szaszi, B., Clelland, H. T., Kovacs, M., Holzmeister, F., van Ravenzwaaij, D., Schulz-Kümpel, H., Hoffmann, S., Nilsson, G., Kosa, L., Torma, Z. A., Abdelfatah, Y., Aberson, C. L., Acar, O. A., Acem, E., Adamkovic, M., Adamovich, T., Adiasto, K., Ahnström, L., ... Nosek, B. A. (2026). Investigating the analytical robustness of the social and behavioural sciences. *Nature*, 652(8108), 135–142. <https://doi.org/10.1038/s41586-025-09844-9>
- Agarwal, J., & Raju, G. (2025). Establishing quality standards for empirical studies using mixed-methods designs in engineering education research. *International Journal of Engineering Education*, 41, 808–820.
- Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 452–454. <https://doi.org/10.1038/533452a>
- Banerjee, R. (2022). Replication research in education: Is the tide turning? *Educational Research and Evaluation*, 27(1–2), 220–224. <https://doi.org/10.1080/13803611.2021.2022311>
- Bartoš, F. (2024). The untrustworthy evidence in dishonesty research. *Meta-Psychology*, 8(1), 1–8. <https://doi.org/10.15626/MP.2023.3987>
- Bartoš, F., & Schimmack, U. (2022). Z-curve 2.0: Estimating Replication Rates and Discovery Rates. *Meta-Psychology*, 6(1). <https://doi.org/10.15626/MP.2021.2720>
- Beddoes, K. (2022). Reflections on the use of theory in engineering education research: Interdisciplinary challenges and comparisons. In S. H. Christensen, A. Buch, E. Conlon, C. Didier, C. Mitcham, & M. Murphy (Eds.), *Engineering, Social Sciences, and the Humanities: Have Their Conversations Come of*

Age? (pp. 179–194). Springer International Publishing.

https://doi.org/10.1007/978-3-031-11601-8_9

Bernhard, J., & Baillie, C. (2016). Standards for quality of research in engineering education. *International Journal of Engineering Education*, 32(6), 2378–2394.

Borrego, M. (2007). Development of engineering education as a rigorous discipline: A study of the publication patterns of four coalitions. *Journal of Engineering Education*, 96(1), 5–18. <https://doi.org/10.1002/j.2168-9830.2007.tb00911.x>

Borrego, M., Douglas, E. P., & Amelink, C. T. (2009). Quantitative, qualitative, and mixed research methods in engineering education. *Journal of Engineering Education*, 98(1), 53–66. <https://doi.org/10.1002/j.2168-9830.2009.tb01005.x>

Brunner, J., & Schimmack, U. (2020). Estimating population mean power under conditions of heterogeneity and selection for significance. *Meta-Psychology*, 4(1), 1–22. <https://doi.org/10.15626/MP.2018.874>

Buckley, J. (2026). Verification studies and cumulative knowledge generation in technology and design education research. *International Journal of Technology and Design Education*. <https://doi.org/10.1007/s10798-026-10082-x>

Buckley, J., Hyland, T., & Seery, N. (2023). Estimating the replicability of technology education research. *International Journal of Technology and Design Education*, 33(4), 1243–1264. <https://doi.org/10.1007/s10798-022-09787-6>

Cai, J. (2022). Promoting conceptual replications in educational research. *Educational Research and Evaluation*, 27(1–2), 8–11. <https://doi.org/10.1080/13803611.2021.2022313>

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Committee on Scientific Principles for Education Research. (2002). *Scientific Research in Education* (R. Shavelson & L. Towne, Eds.). National Research Council, National Academy Press.

<https://www.nationalacademies.org/read/10236/chapter/1>

Epskamp, S., & Nuijten, M. (2018). *statcheck: Extract statistics from articles and recompute p values* (Version 1.3.0.) [RStudio]. <https://CRAN.R-project.org/package=statcheck>

Gupta, A., & Bosco, F. (2023). Tempest in a teacup: An analysis of p-Hacking in organizational research. *PLOS ONE*, 18(2), e0281938.
<https://doi.org/10.1371/journal.pone.0281938>

Loui, M. C., & Borrego, M. (2019). Engineering Education Research. In A. V. Robins & S. A. Fincher (Eds.), *The Cambridge Handbook of Computing Education Research* (pp. 292–322). Cambridge University Press.
<https://doi.org/10.1017/9781108654555.012>

Miske, O., Abatayo, A. L., Daley, M., Dirzo, M., Fox, N., Haber, N., Hahn, K. M., Struhl, M. K., Mawhinney, B., Silverstein, P., Stankov, T., Tyner, A. H., Adamkovič, M., Alzahawi, S., Anafinova, S., Awtrey, E., Axze, E., Bailey, J., Bakker, B. N., ... Errington, T. M. (2026). Investigating the reproducibility of the social and behavioural sciences. *Nature*, 652(8108), 126–134.
<https://doi.org/10.1038/s41586-026-10203-5>

Morrison, K. (2022). Conceptual replications, research, and the “what works” agenda in education. *Educational Research and Evaluation*, 27(1–2), 35–60.
<https://doi.org/10.1080/13803611.2021.2022314>

Nosek, B. A., Allison, D. B., Jamieson, K. H., McNutt, M., Nielsen, A. B., & Wolf, S. M. (2026). A framework for assessing the trustworthiness of scientific

- research findings¹. *Proceedings of the National Academy of Sciences*, 123(6), e2536736123. <https://doi.org/10.1073/pnas.2536736123>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6), 528–530. <https://doi.org/10.1177/1745691612465253>
- Perry, T., Morris, R., & Lea, R. (2022). A decade of replication study in education? A mapping review (2011–2020). *Educational Research and Evaluation*, 27(1–2), 12–34. <https://doi.org/10.1080/13803611.2021.2022315>
- Perry, T., & See, B. H. (2022). Replication study in education. *Educational Research and Evaluation*, 27(1–2), 1–7. <https://doi.org/10.1080/13803611.2021.2022307>
- Pitterson, N., Allendoerfer, C., Streveler, R., Ortega-Alvarez, J., & Smith, K. (2020). The importance of community in fostering change: A qualitative case study of the rigorous research in engineering education (RREE) program. *Studies in Engineering Education*, 1(1). <https://doi.org/10.21061/see.7>
- Power, J. (2021a). Enhancing engineering education through the integration of Open Science principles: A strategic approach to systematic reviews. *Journal of Engineering Education*, 110(3), 509–514. <https://doi.org/10.1002/jee.20413>
- Power, J. (2021b). Systematic reviews in engineering education: A catalyst for change. *European Journal of Engineering Education*, 46(6), 1163–1174. <https://doi.org/10.1080/03043797.2021.1980770>

- Satya Putra, A., Morpew, J., & Tan, L. (2023). By the numbers: A review of quantitative research methods in Journal of Engineering Education from 2012 to 2022. *2023 ASEE Annual Conference & Exposition Proceedings*, 43097. <https://doi.org/10.18260/1-2--43097>
- Schimmack, U. (2022). 2021 Replicability Report for the Psychology Department at the University of Amsterdam. *Replicability-Index*. <https://replicationindex.com/2022/03/29/rr22-u-amsterdam/>
- Schimmack, U. (2025a, July 8). *Z-Curve.3.0 Tutorial: Introduction - Replicability-Index*. Replicability-Index: Improving the Replicability of Empirical Research. <https://replicationindex.com/2025/07/08/z-curve-tutorial-introduction/>
- Schimmack, U. (2025b, July 12). *Z-Curve.3.0 Tutorial: Chapter 2 - Replicability-Index*. Replicability-Index: Improving the Replicability of Empirical Research. <https://replicationindex.com/2025/07/11/z-curve-3-0-tutorial-chapter-2/>
- Schimmack, U., & Bartoš, F. (2023). Estimating the false discovery risk of (randomized) clinical trials in medical journals based on published p-values. *PLOS ONE*, 18(8), e0290084. <https://doi.org/10.1371/journal.pone.0290084>
- Shrout, P., & Rodgers, J. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, 69(1), 487–510. <https://doi.org/10.1146/annurev-psych-122216011845>
- Soto, M. D., & Schimmack, U. (2025). Credibility of results in emotion science: A Z-curve analysis of results in the journals Cognition & Emotion and Emotion. *Cognition and Emotion*, 39(8), 1803–1819. <https://doi.org/10.1080/02699931.2024.2443016>

- Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, 10(2), 220346. <https://doi.org/10.1098/rsos.220346>
- The R Foundation for Statistical Computing. (2022). *R: A language and environment for statistical computing* (Version 4.2.2 “Innocent and Trusting”) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Tyner, A. H., Abatayo, A. L., Daley, M., Field, S., Fox, N., Haber, N. A., Hahn, K. M., Struhl, M. K., Mawhinney, B., Miske, O., Silverstein, P., Soderberg, C. K., Stankov, T., Abbasi, A., Aberson, C. L., Aczel, B., Adamkovič, M., Albayrak, N., Allen, P. J., ... Errington, T. M. (2026). Investigating the replicability of the social and behavioural sciences. *Nature*, 652(8108), 143–150. <https://doi.org/10.1038/s41586-025-10078-y>
- Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, 13(4), 411–417. <https://doi.org/10.1177/1745691617751884>
- William, D. (2022). How should educational research respond to the replication “crisis” in the social sciences? Reflections on the papers in the Special Issue. *Educational Research and Evaluation*, 27(1–2), 208–214. <https://doi.org/10.1080/13803611.2021.2022309>