Examining the Replicability of Contemporary Technology Education Research

Jeffrey Buckley, Tomás Hyland and Niall Seery

Discourse in psychological science concerning the replication crisis is growing, with typically cited indicators including scientific fraud, the use of questionable research practices and the failure of published effects to replicate. It is paramount that education researchers are aware of this crisis and adopt appropriate methodological and reporting practices so as to ensure that a merited degree of trust can be placed in published results. As discourse concerning replicability is currently scarce in technology education, this study presents a z-curve analysis of the replicability of contemporary technology research to instigate this type of consideration. Articles from volumes 27, 28 and 29 of the International Journal of Technology and Design Education were included in the analysis. Results show an increase in the replicability rate from 64% in volume 27 to 70% in volume 28 and finally 71% in volume 29 which should be considered as quite good. However, there is room for improvement to ensure confidence in technology education research, and practices to improve as a field are discussed.

Keywords: Replicability, Statistical power, Research methods, Z-curve analysis.

Introduction

Despite its importance, the concept of *replicability* in research elicits a degree of contention with respect to its definition (National Academies of Sciences Engineering and Medicine, 2019). This is largely due to its similarity with the concept of *reproducibility*. In a review paper concerned with the usage of these terms, Barba (2018) noted three categories of usage, which she termed as groups A, B1, and B2. Researchers in group A used the terms "replicability" and "reproducibility" interchangeably. Those in group B1 used the term "reproducibility" for instances where the original researchers' data and computer codes are used to regenerate the same results and the term "replicability" for when a researcher collects new data which supports the results of a previous study. Finally, group B2 used the term "reproducibility" to refer to when independent researchers arrive at the same conclusions using their own methods and data and the term "replicability" for a different researcher(s) come to the same conclusion using the original researchers source materials and methodology, thereby being in near direct opposition to those in group B1. Therefore, to clarify for the purposes of this article, these terms will be considered in line with their usage by the American National Academies of Sciences, Engineering, and Medicine (2019). Replicability will be used to refer to consistency in results across studies aiming to answer the same scientific question while reproducibility is considered to mean obtaining consistent results when using the same input data.

Replication is critical to building confidence in research results as consistent replications of results suggest validity and, depending on other methodological factors, generalisability. Likewise, when results fail to replicate, evidence is provided to suggest that they may lack validity. At least within psychological science, "until recently, psychologists were confident that published results were replicable" (Brunner & Schimmack, 2017, p. 3). This is due to published studies generally reporting statistically significant results, meta-analyses generally supporting empirical hypotheses, and multi-study articles typically reporting series of successful replications. However, a what has since become referred to as a "replication crisis" arose in the 2010's after a series of events which cast doubt on the replicability of psychological research. These included failed replications of psychological results,

reports of psychologists unwilling to share their data, and evidence that many psychologists admitted to engaging in at least one questionable research practice (Pashler & Wagenmakers, 2012). Unfortunately, while the most desirable explanation for published research results would be that they align with correct theory, there are many other explanations such as questionable research practices, publication bias, statistical error, and lack of consideration of relevant variables.

Replication failures have been observed across a variety of areas of psychology, including educational psychology. This can have significant implications as research is translated into practice such as wasting resources, having students participate in ineffective or even harmful interventions, general advocacy for sub-standard pedagogical methodologies, diminished confidence in research and researchers from the general public and other relevant stakeholders, etc. While to date there have been no notable replication failures in the technology education research, it is critical that researchers within the field consider the replicability of relevant results, engage in replication studies, and stay up to date with contemporary research practices, particularly as many have emerged in light of the replication crisis. By way of example, the popular theory of growth mindset has recently seen an emergence of studies which call both the theory and associated interventions into question. For example, Li and Bates (2017) failed to replicate Mueller and Dweck's (1998) seminar study. Studies by Foliano, Rolfe, Buzzeo, Runge and Wilkinson (2019) and Rienzo, Rolfe and Wilkinson (2015) found teaching students growth mindset to have no impact on grades or other outcomes. Similarly, a meta-analysis (Sisk, Burgoyne, Sun, Butler, & Macnamara, 2018) and large scale study (Bahník & Vranka, 2017) found the correlation between growth mindset and academic performance to be negligible. These results cast doubt over the validity of the theory and utility of pertinent interventions. Engaging in quality control research activities such as examining the replicability of technology education research could mitigate the occurrence of a similar event within the field. This paper reports the results of an initial investigation into the replicability of recently published results from within technology education in an attempt to instigate discourse associated with enhancing relevant research practices.

Method

Statistical approach: Z-curve analysis

There are multiple approaches to estimating the replicability of studies. For example, the Many Labs approach is to replicate a single study across several labs (Klein et al., 2014). This approach requires significant resources. There are also statistical methods, such as p-curve and z-curve analyses. These approaches are useful for estimating the replicability of a large number of original studies by estimating the average statistical power of studies which produced a statistically significant result. Brunner and Schimmack (2017) recommend the use of z-curve over p-curve as while they found p-curve to perform well when studies were homogeneous, it did not where there was heterogeneity. Z-curve however was designed to estimate power under the condition that population effect sizes are heterogeneous.

This study used z-curve 2.0. (Bartoš & Schimmack, 2020), an extension of z-curve 1.0. (Bartoš & Schimmack, 2020). The evolution of z-curve up to the current version 2.0. (at the time of publication of this manuscript) is described in detail by Brunner and Schimmack (2016, 2017, 2020) and Bartoš and Schimmack (2020). These articles include simulation studies, comparisons with other methods, descriptions of associated formulae, and descriptions and rationales for the z-curve approach. For this article therefore it is more pertinent to focus on the outputs it provides. These include:

- Observed discovery rate (ODR): This is the proportion of statistically significant results observed within the dataset. For example, if the dataset contained 100 statistical tests, and 70 of them provided a statistically significant ($p < .05$) result, the ODR would be .70.
- Expected discovery rate (EDR): Based on the significant results reported, Z-curve 2.0. has the capacity to estimate the total number of studies that were conducted, including those with non-

significant results which may not have been reported (Bartoš & Schimmack, 2020). This allows for and estimation of the EDR, i.e., the percentage of significant results obtained across all studies. The EDR can be compared with the ODR with discrepancies indicating publication bias.

- File drawer ratio (FDR): The FDR is an estimation of how many non-significant results were obtained for each significant result. It is a conversion of the EDR, i.e., FDR = (1-EDR)/EDR. The idea of a file drawer was coined by Rosenthal (1979) where he described metaphorical "file drawers" in research labs filled with non-significant results which were not reported.
- Soric's false discovery risk (SFDR): The SFDR is the maximum false discovery rate (Soric, 1989). For example, an SFDR value of .08 or 8% indicates that no more that 8% of the significant results from the z-curve analysis are false discoveries. The SFDR is calculated as a function of the FDR, i.e., SFDR = FDR*alpha/(1-alpha) or SFDR = FDR*0.5/.95.
- Expected replicability rate (ERR): The ERR refers to the unconditional mean power of the studies within the dataset with statistically significant results, as this can be used to predict the outcome of exact replication studies.

Dataset

As the purpose of this study was to estimate the replicability of recently published technology education research, predominantly as a conduit for introducing discourse concerning replication into technology education, articles from 2017-2019 (volumes 27, 28, and 29) from the *International Journal of Technology and Design Education* (IJTDE) were selected as the dataset. As the input for z-curve are *z*-scores derived from test statistics and *p*-values, only quantitative studies which reported statistical tests e.g., *t*, *F*, chi-squared, or *z* statistics, were eligible.

Each article from volumes 27 (n = 38), 28 (n = 54), and 29 (n = 59) was manually screened and coded initially as being empirical or non-empirical. Then empirical studies were coded as being either quantitative, qualitative, or mixed methods based on the nature of the empirical work which was reported. Finally, quantitative and mixed methods studies were coded as to whether they did or did not report the results of a hypothesis test. Figure 1 illustrates the results of this coding process across each of the three volumes, and Figure 2 illustrates a breakdown across each issue. There is a relatively large variation with respect to the nature of published work across issues as can be expected, but the nature of work published each year is relatively stable.

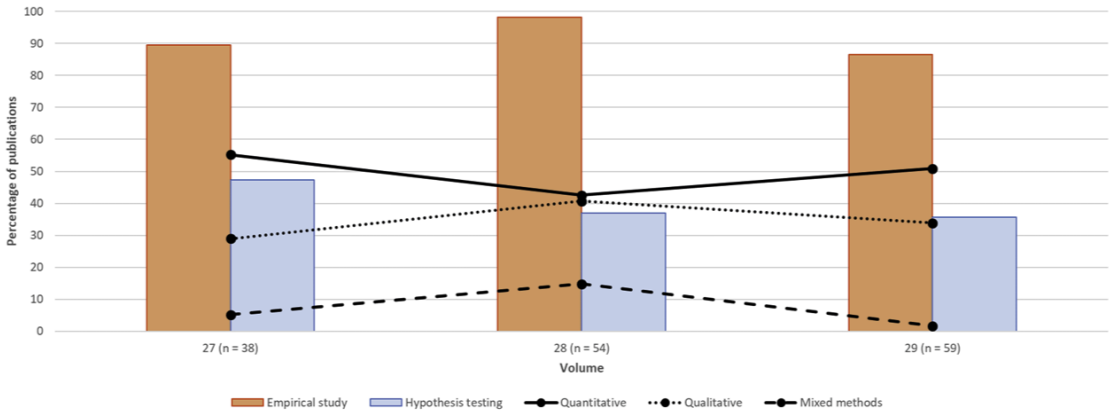

Figure 1. Descriptive statistics of IJTDE articles illustrating the percentage of published articles which were empirical, included hypothesis testing, and which reported on quantitative, qualitative, and mixed methods studies, across volumes 27 to 29.

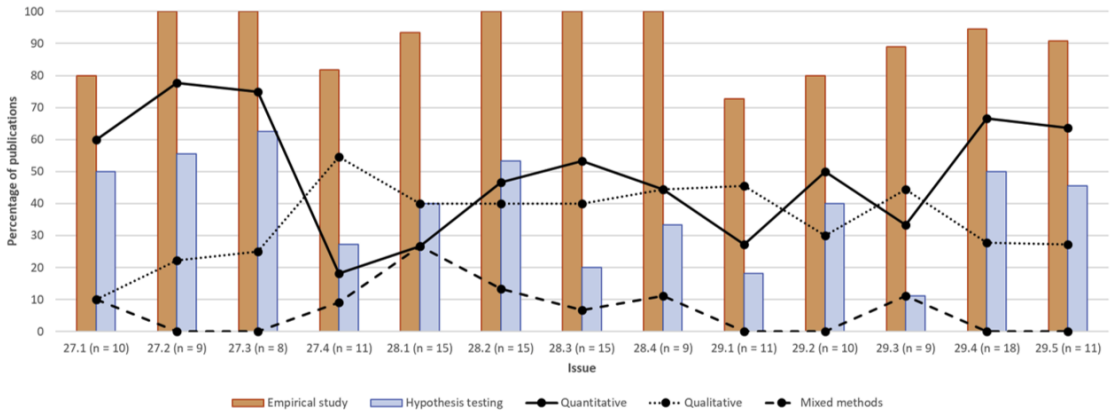

Figure 2. Descriptive statistics of IJTDE articles illustrating the percentage of published articles which were empirical, included hypothesis testing, and which reported on quantitative, qualitative, and mixed methods studies, across issues 27.1 to 29.5.

Figure 2 represent all articles published across volumes 27, 28 and 29 of the IJTDE, however the z-curve analysis was conducted between the publishing of issues 29.4 and 29.5. Therefore, the z-curve analyses are only associated with issues 27.1 to 29.4. Across these issues, a total of 54 articles reported tests statistics associated with a hypothesis test, with most articles reporting multiple test statistics. Some articles did not provide sufficient information, for example reporting a statistically significant mean difference but only presenting $p < .05$, and thus could not be coded. In total, there were 18 articles in volume 27 which reported on hypothesis testing providing a total of 233 valid test statistics, there were 20 pertinent articles in volume 28 which provided a total of 279 valid test statistics, and there were 16 pertinent articles in volume 29 (issues 29.1 to 29.4) which reported a total of 178 valid test statistics.

Each test statistic was manually extracted and exact *p* values were computed if they were not reported which were then converted to *z*-scores. In a two-tailed test for example, a *p*-value of 0.05 equates to a *z*-score of 1.96. A separate dataset for each volume of the IJTDE was then created, which consisted of a single list of the valid *z*-scores derived from the reported test statistics in that volume. Finally, the z-curve analysis was conducted in RStudio using the *z*-curve package developed by František (see Schimmack, 2020). At this point it should be noted that, as quite often no focal test was apparent, all reported test statistics were included in the analysis. Kalmendal and Mühlmeister (2019) compared the use of z-curve with just focal test statistics and with all tests statistics after automated extraction and found that automated extraction was as good as manual coding, suggesting the inclusion of all test statistics as was done in this study is not a significant limitation.

Results

The results of the z-curve analysis for volume 27 are presented in Figure 3. There is a significant difference (the confidence intervals do not overlap) between the ODR and EDR suggesting the presence of publication bias. The FDR of 4.16 suggests that for each significant result reported, there are approximately 4.16 non-significant results which were not reported. The SFDR value of .22 suggests that a maximum of 22% of the 132 significant results were false discoveries. Finally, the ERR of .64 suggests a replicability rate of 64% for the reported significant results in that volume.

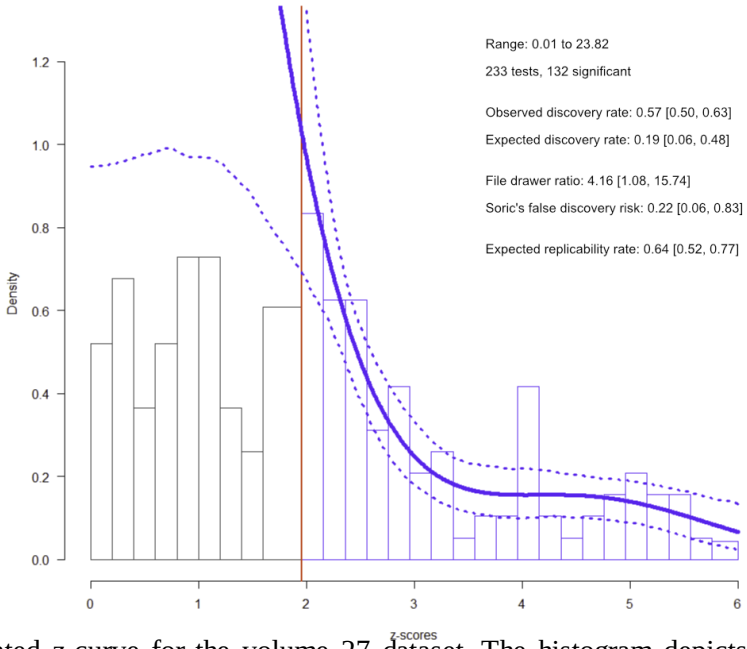

Figure 3. Estimated z-curve for the volume 27 dataset. The histogram depicts the distribution of z-scores. The vertical red line represents the statistical significance criterion (z = 1.96/p = 0.05). The solid blue line displays the density of the estimated model. The dotted lines represent uncorrected piece-wise 95% confidence intervals.

The results of the z-curve analysis for volume 28 are presented in Figure 4. There is a significant difference (the confidence intervals do not overlap) between the ODR and EDR suggesting the presence of publication bias. The FDR of 2.57 suggests that for each significant result reported, there are approximately 2.57 non-significant results which were not reported. The maximum SFDR value of .14 suggests that a maximum of 14% of the 191 significant results were false discoveries. Finally, the ERR of .70 suggests a replicability rate of 70% for the reported significant results in that volume.

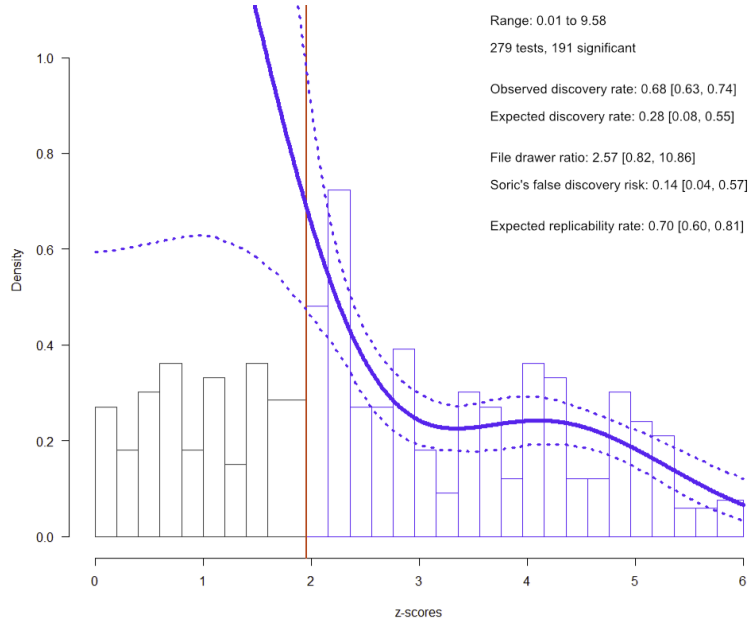

Figure 4. Estimated z-curve for the volume 28 dataset. The histogram depicts the distribution of z-scores. The vertical red line represents the statistical significance criterion (z = 1.96/p = 0.05). The solid blue line displays the density of the estimated model. The dotted lines represent uncorrected piece-wise 95% confidence intervals.

The results of the z-curve analysis for volume 29 are presented in Figure 5. The overlap in the confidence intervals between the ODR and the EDR indicates there is not a statistically significant difference between the two, and the difference is less that what was observed for volumes 27 and 28, indicating that while there may be publication bias, at least in this volume it was lower than the previous two. The FRD of 2.40 suggests that for each significant result reported, there are approximately 2.40 non-significant results which were not reported. The maximum SFDR value of .13 suggests that a maximum of 13% of the 88 significant results were false discoveries. Finally, the ERR of .71 suggests a replicability rate of 71% for the reported significant results in that volume.

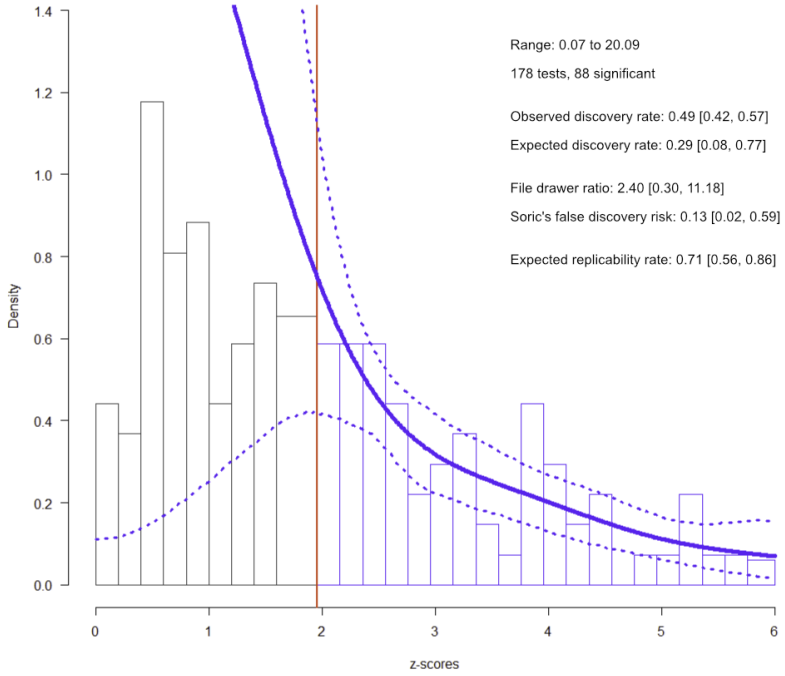

Figure 5. Estimated z-curve for the volume 29 dataset. The histogram depicts the distribution of z-scores. The vertical red line represents the statistical significance criterion (z = 1.96/p = 0.05). The solid blue line displays the density of the estimated model. The dotted lines represent uncorrected piece-wise 95% confidence intervals.

Discussion

Before reflecting on these results, it is first important to consider the concepts of statistical power and statistical errors. Statistical power describes the probability of detecting an effect if there is an effect there to be detected. A general recommendation for the amount of statistical power which should be planned for is 80% (Cohen, 1988). With this level of power, this would mean that if an original study found a true, statistically significant effect, if researchers were to conduct ten replication studies of this, 8 would produce a significant result. If there is a true effect, and a study fails to identify it, a type II error is committed. Similarly, if there is no effect to be found, and a study observes a statistically significant result, a false positive or type I error is committed. While an 80% recommendation is often cited for statistical power (the probability of avoiding a type II error), a 5% recommendation is typically used as an alpha value (the probability of avoiding a type I error). This is often denoted in quantitative studies as $p < .05$.

Statistical power is directly related to a studies sample size and the effect size of the phenoma being investigated. Consider a study with a very large or obvious effect, there is not a need to have as many participants as would be needed if the effect being investigated is very small. Therefore, the larger the

effect size, the smaller the required sample size is to achieve 80% power. In practice, for researchers this means reviewing literature of similar studies to estimate a hypothesised effect size or identifying the smallest effect size of interest, and conducting a power analysis to determine the required sample size to achieve the desired level of statistical power. For researchers trying to avoid committing a type I error, adjustment of the alpha level used can be considered. For example, if a study is conducted where one statistical test is needed and reported on, using an alpha level of .05 means that there is a 5% probability that if the result is statistically significant it was a false positive, i.e. a 1 in 20 liklihood. However if a study includes more statistical tests (multiple comparisons) using the same dependant variable, each measured against an alpha value of .05, the family-wise error rate increses (Figure 6). Researchers can adjust for multiple comparisons by, for example, using corrections such as the Bonferroni or Holm corrections.

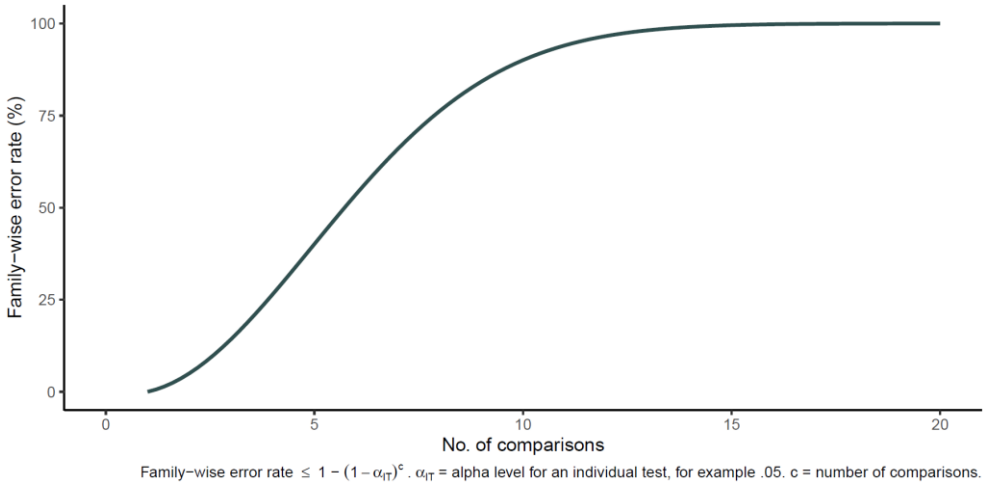

Figure 6. Family-wise error rate as a function of the number of statistical comparisons made in a study.

From the z-curve analysis, the EDR value describes the unconditional mean power of all studies within a dataset, including those with nonsignificant results, while the ERR value describes the unconditional mean power of the studies with significant results. There are different amounts of power between these values as publication bias (selection for statistical significance) favours studies with higher power as they are more likley to produce significant results. In the event that all studies within a dataset were designed to have the same power however these values would be identical. As there is a recommendation to design studies to have 80% power, values lower than this indicate that within a dataset they studies are generally underpowered, thus increasing the liklihood of committing a type II error and finding a false negative.

From the results of this analysis, the ERR values increased across each volume from 64% in volume 27 to 71% in volume 29 (issues 29.1 to 29.4). This indicates an increase in the replicability of research being published in the IJTDE. However, these results are likley an overestimation as all test statistics were included whether they were meaningful hypotheses or not. So for example, if a full correlation matrix was included, each correlation was considered within the analysis when the authors and research question may only have been interested in a few. That said, replicability rates of near 70% should be considered quite good considering the general rule of aiming for 80% statistical power, but considering this is likley an overestimation, there is still room for improvement in technology education research.

In terms of implications for practice, an observation from the screening process at the beginning of this study was that many articles do not clearly report the results of statistical tests or do not report complete information. Degrees of freedom for example were often ommitted, as were test statistics and effect sizes. Not including all relevant information makes it difficult and sometimes impossible to evaluate the

results. Technology education research would also benefit from making considerations of replicability more widespread. Again, based on observations from reviewing the articles analysed for this article, considerations of sample size through conducting power analyses is not a practice typical of technology education research. Similarly, there was little to no evidence of consideration for multple comparisons. A fourth type of activity which would benefit technology education research would be to conduct more replication studies of theoretically interesting results to aid in determining the validity of results. In conjunction, if technology education journals were to introduce the option to publish registered reports (submiting a study plan to a journal which is peer reviewed prior to conducting a study, and upon receiving conditional acceptance conducting the study which, even if the result is not statistically significant, is accepted provided it is well reported and aligns with the study plan), issues associated with publication bias could be mitigated.

References

Bahník, Š., & Vranka, M. (2017). Growth mindset is not associated with scholastic aptitude in a large sample of university applicants. *Personality and Individual Differences*, *117*(1), 139–143.

Barba, L. (2018). Terminologies for reproducible research. *ArXiv*. https://arxiv.org/pdf/1802.03311

Bartoš, F., & Schimmack, U. (2020). Z-curve.2.0: Estimating replication rates and discovery rates. *PsyArXiv*. https://doi.org/10.31234/osf.io/urgtn

Brunner, J., & Schimmack, U. (2016). *How replicable is psychology? A comparison of four methods of estimating replicability on the basis of test statistics in original studies*. http://www.utstat.utoronto.ca/~brunner/papers/HowReplicable.pdf

Brunner, J., & Schimmack, U. (2017). Z-curve: A method for the estimating replicability based on test statistics in original studies. *OSF Preprints*. https://doi.org/10.31219/osf.io/wr93f

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

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

Foliano, F., Rolfe, H., Buzzeo, J., Runge, J., & Wilkinson, D. (2019). *Changing mindsets: Effectiveness trial*. London: Educational Endowment Foundation, National Institute of Economics and Social Research.

Kalmendal, A., & Mühlmeister, T. (2019). *Predicting the replicability of experimental research in work and organizational psychology published in the Journal of Applied Psychology* (Master's Thesis). Linnaeus University, Sweden.

Klein, R., Ratliff, K., Vianello, M., Adams Jr., R., Bahník, Š., Bernstein, M., … Nosek, B. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152.

Li, Y., & Bates, T. (2017). Does growth mindset improve children's IQ, educational attainment or response to setbacks? *SocArXiv*. https://doi.org/10.31235/osf.io/tsdwy

Mueller, C., & Dweck, C. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology*, *75*(1), 33–52.

National Academies of Sciences Engineering and Medicine. (2019). *Reproducibility and replicability in science*. Washington, DC: The National Academies Press.

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.

Rienzo, C., Rolfe, H., & Wilkinson, D. (2015). *Changing mindsets: Evaluation report and executive summary*. London: Educational Endowment Foundation, National Institute of Economics and Social Research.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641.

Schimmack, U. (2020). Z-Curve.2.0. Retrieved January 13, 2020, from https://replicationindex.com/2020/01/10/z-curve-2-0/

Sisk, V., Burgoyne, A., Sun, J., Butler, J., & Macnamara, B. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science*, *29*(4), 549–571.

Soric, B. (1989). Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*, *84*(406), 608–610.

Dr *Jeffrey Buckley* is a researcher in technology education with specific interest in how people learn, and research methods and practices. He is a Lecturer at Athlone Institute of Technology, Ireland, and an Affiliate Faculty Member of the Department of Learning at KTH, Royal Institute of Technology, Sweden. He is a member of the Technology Education Research Group (TERG) and the Learning in Engineering Education and Progress (LEEaP) research group.

Tomás Hyland is PhD student at Athlone Institute of Technology, Ireland, where he is studying the relationship between spatial ability and learning in technology and engineering education. He is particularly interested in classroom-based research. He is a member of the Technology Education Research Group (TERG).

Dr *Niall Seery* is currently Deputy President of Athlone Institute of Technology, Ireland. He served as Vice President of Academic Affairs and Registrar before taking the role as Director of the Technological University Project at AIT. He has a PhD in Engineering Education and in 2010 he founded and continues to direct the Technology Education Research Group (TERG), where he is still active in research development and mentorship.