False discoveries and erroneous findings in randomised controlled trials in education

Sam Sims, UCL Institute of Educaiton
Jake Anders, UCL Institute of Education
Matthew Inglis, Loughborough University
Hugues Lortie-Forgues, University of York

Motivation

Randomised controlled trials (RCTs) have proliferated in education research in recent years. Funders are attracted to descriptions of a "gold standard" (Pocock, 1982) of evidence as part of a wider "what works" approach to policy. This increased focus partly emerges from strong advocacy for increased use of RCTs from voices within government and from other fields of research where their use is more prevalent (e.g. Haynes et al., 2012). Such actors have typically drawn attention to the need to challenge overconfidence in perceptions that educators already know whether what we are doing is productive. In England, this approach has been institutionalised through the establishment of a network of "what works" centres, where the what works centre for education is the Education Endowment Foundation (EEF).

Meanwhile, researchers are drawn to RCTs by the promise of an unbiased estimator of the causal impact of interventions. This is otherwise extremely difficult to obtain in the evaluation of such interventions because there are typically a multitude of measurable and unmeasurable factors that confound our observational estimators of impact based on comparing those who self-select into obtaining treatment with those who choose not to do so. This can be shown more formally with Rubin's (1974) potential outcomes framework. RCTs attempt to solve this issue by using randomisation to replace this self-selection, breaking any systematic correlation between receiving treatment and the characteristics of those who do so or do not.

However, it is increasingly recognised that many education trials are not sufficiently statistically powered to detect effects of the size typically estimated in the literature (Lortie-Forgues & Inglis, 2019), if indeed there is an effect at all. This leads to trials that have been labelled "uninformative" in that we haven't found a statistically significant effect, but nor can we be particularly confident in the null hypothesis of no effect.

This paper draws attention to a set of closely related issues. However, rather than focusing on when trials provide little or no information, we focus instead on when they provide incorrect information. More specifically, we ask: when we conclude based on an RCT that there is an effect, what is the probability there is in fact no effect (a false discovery), the true effect is in the opposite direction (Type-S error) or the true effect is in the same direction but is of a different magnitude (Type-M error) (Colquhoun, 2019; Gelman & Carlin, 2014).

Methods

We begin by setting out a unified conceptual framework for understanding all three types of errors, grounded in potential outcomes notation and an adapted version of the diagrams developed by Gelman & Tuerlinckx (2002). In order to assess the likely incidence and severity of these phenomena, we then calculate empirical estimates of all three types of error using a set of thirteen completed trials deemed to have delivered ‘promising’ results by the Education Endowment Foundation. Our approach is therefore a retrospective design analysis (Gelman & Carlin, 2014).

For a given trial, the probability of a false positive conclusion depends on the Bayes Factor (the ratio of how well the alternative and null hypothesis predict the data) and the prior odds of the intervention being effective (Colquhoun, 2019). Similarly, the probability of Type-S and Type-M error depends on the observed effect, standard error and p-value, as well as a prior about the true effect size. We derive our priors about the odds of an intervention being effective or the size of the true effect from separate, out-of-sample studies, largely drawn from the National Centre for Education Evaluation (NCEE) database of trials. In addition, we check for the stability of estimates under a range methods for calculating these priors, namely: the average NCEE effect size; the subject-specific NCEE effect size and an assessment based on the literature from non-NCEE evaluations of very similar interventions, in the spirit of those reported in Gelman & Carlin (2014).

Contribution

This paper aims to draw attention to the wider issues raised by lack of power in educational trials beyond the potential to produce findings that are ‘uninformative’ in the way outlined by Lortie-Forgues & Inglis (2019). We highlight the implications of these results and propose steps that should be taken by researchers and funders to mitigate these implications in the interpretation of results and funding decisions that result from these.

Our current (provisional) results suggests that:
- the probability of a false positive varies widely across trials, between 0.02 and 0.44
- the probability of a Type-S error is low, between 0.001-0.1
- the magnitude of Type-M error varies widely across trials, ranging from 1.001 to 7.85

The additional information provided by our post-hoc design analysis sheds new light on the amount of information contained in RCTs deemed to have provided ‘promising’ evidence for education interventions. In particular, several of these trials appear to have a non-trivial risk of being false positives and/or providing exaggerated effect size estimates.

We advocate that researchers should consider using such design analysis when planning and designing RCTs in education, as well as post-hoc design analysis when reporting and interpreting the results of trials. In addition, we argue that those responsible for commissioning follow up evaluations should carefully consider the results of post-hoc design analysis–rather than only relying on effect size and p-values–when deciding whether to fund scale-up or effectiveness trials.

Bibliography

Colquhoun, D. (2019). The false positive risk: A proposal concerning what to do about p-values. *The American Statistician*, *73*(1), 192-201.
Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspectives on Psychological Science, 9(6), 641–651. https://doi.org/10.1177/1745691614551642
Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, *15*(3), 373-390.

Haynes, L., Service, O., Goldacre, B. & Torgerson, D. (2012). Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials. London: UK Cabinet Office.
Lortie-Forgues, H., & Inglis, M. (2019). Rigorous Large-Scale Educational RCTs Are Often Uninformative: Should We Be Concerned? Educational Researcher, 48(3), 158–166. https://doi.org/10.3102/0013189X19832850
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688-701. http://dx.doi.org/10.1037/h0037350
Pocock, S. J. (1982), Statistical Aspects of Clinical Trial Design†. Journal of the Royal Statistical Society: Series D (The Statistician), 31: 1-18. doi:10.2307/2988097