Quantifying 'promising trials bias' in randomized controlled trials in education

Sam Sims, Jake Anders, Matthew Inglis, Hugues Lortie-Forgues

Journal of Research on Educational Effectiveness

Randomized controlled trials (RCTs) have proliferated in education, in part because they provide an unbiased estimator for the causal impact of interventions. Yet RCTs are only unbiased in expectation (on average across many RCTs).

Estimates of the effect size from specific RCTs will in general diverge from the true effect due to chance differences between the treatment and control group. In suitably powered trials, this imbalance tends to be small and statistical inference helps to control erroneous findings.

Promising Trials Bias

Crucially, however, among the RCTs deemed to show promising results based on meeting a statistical significance threshold (e.g., $p<0.05$) we would expect this random error to systematically inflate effect size estimates. We refer to this as *promising trials bias*.

To see why, consider that when researchers set a threshold for statistical significance of $p<0.05$ an estimate must be 1.96 standard errors away from zero to be declared a promising result. The most exaggerated estimates of effect size are systematically more likely to clear this threshold.

This is particularly problematic in a trial with lower power, where the need to be 1.96 standard errors away from zero represents a higher threshold, thus filtering out all but the most exaggerated estimates. Many RCTs in education have low power (Spybrook, Shi, & Kelcey, 2016; Lortie-Forgues & Inglis, 2019).

Quantifying Promising Trials Bias

How large is promising trials bias in practice? Recall that RCTs are unbiased on average across many studies but, when we focus on a specific estimate from a single trial, chance imbalance means that the estimate will contain error of some magnitude. The retrospective design analysis that we conduct in this paper aims to move back in the other direction (from the specific trial to the general) by asking: for each published RCT, what results would we be likely to obtain under hypothetical replications of the study? We apply this method to 23 RCTs from the UK Education Endowment Foundation archive deemed to have shown promising results. Retrospective design analysis requires us to make assumptions about the true effect size and we calibrate these using three different sets of empirical evidence.

Depending on which of our three assumptions we use, our estimates of promising trials bias range from 1.52 (suggesting effect size estimates are exaggerated by 52%) up to 5.55. While these results are clearly sensitive to the specific assumptions selected, all three sets of assumptions suggest substantively important inflation of estimates.

Implications

Our findings have three practical implications. First, educators and policymakers looking to understand the benefits of different programs should expect smaller effects than suggested by trial results listed in 'warehouses' or 'toolkits' of promising interventions.

Second, researchers should consider applying design analysis prospectively (when designing trials) and retrospectively (when analyzing trials). Gelman and Carlin (2014) describe this as moving from asking only "What is the power of a test?" to also asking "What might be expected to happen in [future] studies of this size?"

Third, researchers should not assume that programs fail to scale up successfully because of the difficulties of implementing at scale. Since education RCTs tend to have low power, we would expect smaller – perhaps much smaller – effects in subsequent effectiveness trials, even if implementation fidelity were perfectly maintained.

Full Article Citation:
Sims, S., Anders, J., Inglis, M., Lortie-Forgues, H. (2022). Quantifying 'promising trials bias' in randomized controlled trials in education. *Journal of Research on Educational Effectiveness.* https://doi.org/10.1080/19345747.2022.2090470.