On the Practicality of Extremely Large Educational RCTs

Hugues Lortie-Forgues and Matthew Inglis

In his response to our article, Sims suggests that increasing the size of educational randomized control trials (RCTs) is a feasible and affordable way to increase their informativeness. We doubt this.

**Keywords:** effect size; evaluation; program evaluation; testing

In our article “Rigorous Large-Scale Educational RCTs Are Often Uninformative: Should We Be Concerned?” we demonstrated that a large proportion of contemporary randomized control trials (RCTs) are not able to detect the small effects typically produced by educational interventions. We also noted that—given how much larger trials would need to be—simply increasing their size seems an impractical solution to this problem. Sims suggests that we should have focused on the absolute weighted average of trials’ effect sizes, rather than their simple weighted average. Based on his reanalysis, Sims concludes that educational trials could affordably be powered to detect effects of 0.08 standard deviations (SDs). He estimates the cost of such a trial to be around $3 million, four to five times the cost of current trials.

We agree with Sims that the absolute weighted average is a more useful figure to consider in this context and are grateful to him for pointing this out. However, we doubt the implications he draws. There are three main reasons: (a) Sims’s figure for the mean absolute effect size is considerably higher than our estimate, (b) trials’ minimum detectable effect sizes (MDESs) should be lower than their mean absolute effect size, and (c) Sims’s cost estimates rest on the contentious assumption that a trial’s MDES and its cost are linearly related. We discuss each point in turn.

**Sims’s Mean Absolute Effect Size Is Too High**

Computing the weighted average absolute effect size from our data—available here: https://doi.org/10.6084/m9.figshare.c.4421087—produces a considerably lower estimate than Sims’s: 0.06 SDs rather than 0.08 SDs (a nontrivial difference: a properly powered independent-samples t-test [alpha: 0.05; power: 80%] requires 4,908 participants to detect an effect of 0.08 SDs, but 8,724 participants to detect an effect of 0.06 SDs). We suspect that this discrepancy is due to how trials with multiple outcomes (around half the trials in our dataset) should be treated. To ensure a representative estimate and avoid violating statistical independence, our analysis is limited to a single randomly chosen outcome per trial, a common procedure (e.g., Lipsey & Wilson, 2001). In contrast, Sims prioritized mathematics-related outcome when available, on the grounds that they tend to be measured more precisely. We question this decision: Prioritizing precisely measured outcomes biases the effect size estimate upward, since, for the same effect, a less noisy measure will typically yield a larger standardized effect size. Moreover, mathematics outcomes are relatively uncommon (present in 27% of trials in our dataset, compared to a figure of 63% for language-related outcomes), making Sims’s estimate somewhat unrepresentative of effects typically observed in educational trials.

**Trials’ MDESs Should Be Lower Than Their Mean Absolute Effect Sizes**

Perhaps more importantly, even if Sims’s estimate were accurate, trials’ MDESs should be lower than the mean absolute effect size across all trials. One reason is that educational trials are heterogeneous; they involve different interventions, age groups, and outcome measures (this is the rationale for treating trials as a random factor in ours and Sims’s analyses). Given this, powering a trial to
the average effect size would mean appropriately powering only a subset of trials—those whose true effects are equal to or above this average. In other words, unless trials’ MDESs are set to be below their expected effect sizes, a large number of them will be underpowered (McShane & Böckenholt, 2014).

A further problem is that larger trials are likely to yield smaller effects. Sims suggests that to be properly powered, trials would require 280 schools, considerably more than existing norms (the median number of schools in Educational Endowment Foundation [EEF] trials is 50). Should we expect these larger trials to generate similar effects to what is currently typical? Larger trials often cannot be implemented with the same scrutiny and fidelity as smaller trials, which increases standard deviations and lowers effect sizes (e.g., Cheung & Slavin, 2016; Weisburd et al., 1993). For this reason, we should expect larger trials to produce smaller effects than those found in smaller trials, again suggesting that Sims’s MDES of 0.08 SDs is somewhat high.

The Relationship Between Trial Cost and MDES Is Nonlinear

Sims used existing trial data to estimate the cost of a trial with an MDES of 0.08 SDs. As he acknowledges, this prediction falls worryingly far outside the data range (see Figure 1 in Sims’s comment). Equally problematic, however, is the assumption that MDES and cost are linearly related. We doubt that this assumption is reasonable. Assuming linearity would mean, for example, that the cost of decreasing a trial’s MDES from 0.25 SDs to 0.20 SDs would be the same as decreasing it from 0.10 SDs to 0.05 SDs. This is surely wrong. The two scenarios require radically different numbers of additional participants (for an independent-samples t-test, 142 and 4,710 extra participants, respectively). The true relationship most likely mirrors that of MDES and sample size, where the cost of reducing a trial’s MDES by a fixed amount becomes increasingly large. If this is correct, then running trials powered to detect an effect size of 0.08 SDs would likely cost more than $3 million.

We are grateful to Sims for his comment and for provoking us to think about these issues further. However, we continue to believe that increasing trials’ sample sizes is unlikely to be an efficient way to reduce the worryingly large proportion of trials that are uninformative. Less costly means of increasing the power of trials, such as using more targeted outcome measures or focusing on more targeted subgroups, appear more promising.

REFERENCES
Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45*(5), 283–292.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: SAGE.
McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, 9(6), 612–625.
Sims, S. (2020). Informing better trial design: A technical comment on Lortie-Forgues and Inglis (2019). *Educational Researcher, 49*(4), 289–290.
Weisburd, D., Petrosino, A., & Mason, G. (1993). Design sensitivity in criminal justice experiments. *Crime and Justice, 17*, 337–379.

AUTHORS
HUGUES LORTIE-FORGUES is a senior lecturer in the Centre for Mathematical Cognition at Loughborough University, Leicestershire, LE11 3TU, United Kingdom; H.Lortie-Forgues@lboro.ac.uk. His research focuses on mathematics education and the evaluation of educational interventions.
MATTHEW INGLIS is a professor of mathematical cognition at Loughborough University, Leicestershire, LE11 3TU, United Kingdom; m.j.inglis@lboro.ac.uk. His research focuses on understanding the cognitive processes involved in mathematical thinking and reasoning.

Manuscript received March 23, 2020
Accepted April 6, 2020