Informing better trial design: A Technical Comment on Lortie-Forgues and Inglis (2019)

Randomised controlled trials (RCTs) balance both the observable and unobservable characteristics of treatment and control groups, while employing minimal assumptions. They are therefore well placed to determine impact. Despite this, Lortie-Forgues and Inglis (2019) (henceforth LF&I) reach the startling conclusion that 40% of education RCTs are uninformative. The Education Endowment Foundation (EFF) - a large trial commissioner - has responded with a defence of their approach (EEF, 2019). In this note, I argue that both LF&I and EEF have missed a feasible, affordable way of improving the informativeness of trials: making them larger.

LF&I review 141 RCTs commissioned by EEF and the National Centre for Educational Evaluation (NCEE). They define a trial as uninformative if the findings are consistent with the intervention being both effective and ineffective. This is operationalised with a Bayes factor, which is the ratio of how well the alternative and null hypotheses predict the data. Following convention, LF&I deem a trial informative if the data is predicted by one hypothesis three times better than the other, and find 40% are uninformative on this basis. Put another way, LF&I find that 93% of trials estimated effect sizes below their minimum detectable effect size (MDES).

LF&I make two recommendations for trial design. First, target interventions at specific groups likely to be most responsive. Second, use proximal outcome measures likely to show a stronger response. Importantly, LF&I are sceptical that trials can be made more informative by increasing sample size, calculating that for an independent samples t-test to detect an effect size of 0.04 (their weighted mean effect size), an individually-randomised trial would require 20,000 participants!

EEF do not dispute the data or analysis used by LF&I - both of which, to the authors credit, are publicly available. Instead, they emphasise that their trials do now include proximal, non-achievement outcomes. While pointing out that their trials have been increasing in size, EEF also doubt that larger trials are the solution. They calculate that a cluster RCT capable of detecting an effect size of 0.05

would require 800 schools (assuming ICC=0.1 and R^2=0.5). EEF argue that this would drastically reduce the number of trials they could run, rendering their overall output less informative.

LF&I and EEF base their sample size calculations on the expected effect size. This is reasonable, given their aim is to determine whether larger trials could detect the effects typically found in the literature. The contention of this article however, is that they miscalculate the expected effect size. Education is an applied field, which aims to identify both effective practices for practitioners to adopt, and ineffective practices, which practitioners should avoid. It follows that the relevant benchmark is actually the expected *absolute* effect size. Using the LF&I data, I calculate this to be 0.08, twice the mean weighted effect size.[1] This has substantial implications for trial design. Indeed, reproducing the EEF sample size calculation using the 0.08 benchmark shows that such a trial would require only 280 schools (140 in treatment).

To quantify the trade-offs, I model the cost of achieving a given MDES. Cost is partly determined by the level of randomisation and type of intervention. I therefore augment the LF&I data with information from EEF: the type of intervention, as coded by EEF (e.g. behaviour) and additional indicators for texting interventions as well as individual versus cluster randomisation. I drop NCEE trials, due to lack of comparable data. Figure 1 plots MDES against cost, overlaid with a regression line showing the predicted relationship, conditional on the additional EEF covariates. The line reaches the target MDES of 0.08 at around £2,500,000 – four times the current average cost (£510,083). This finding relies on extrapolation along the x axis, but this by design - we need future trials to be different to past trials.

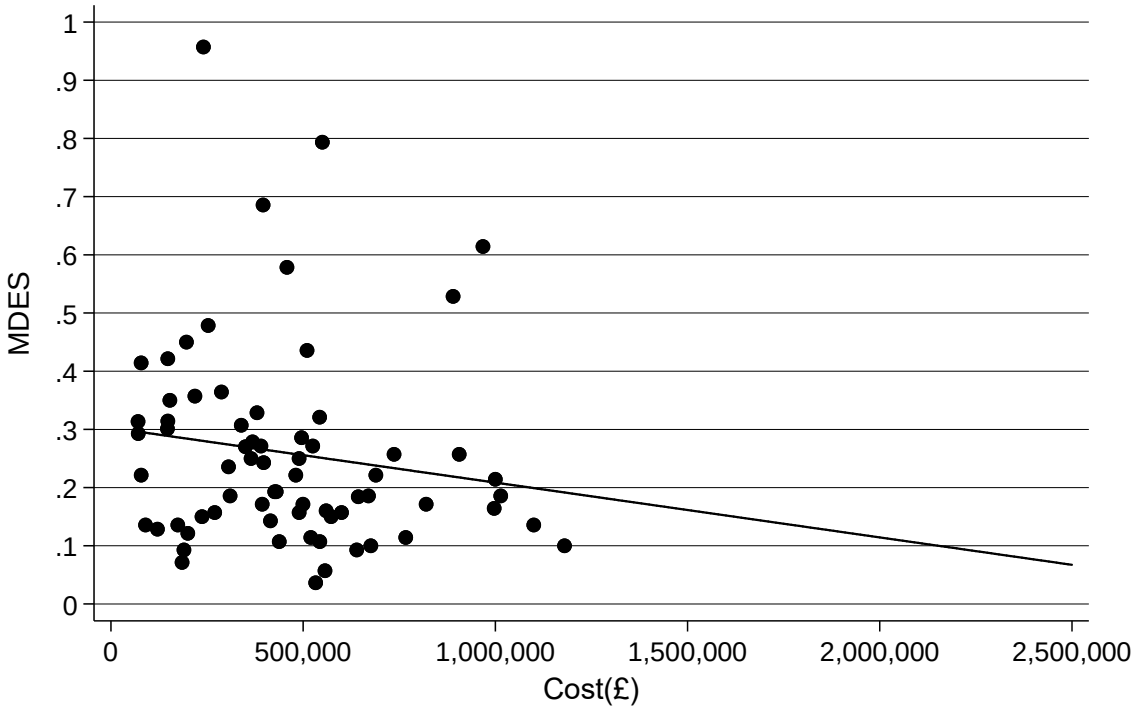

Figure 1. Scatterplot of minimum detectable effect size (MDES) and intervention cost. N=69.

LF&I demonstrate convincingly that trials are often uninformative with regards to impact. Their findings are important and deserve to prompt a rethink of trial design. However, both LF&I and EEF have been too quick to dismiss the value of making trials larger, due to miscalculating the expected effect size. Trials with MDES of 0.08 are affordable, requiring cluster RCTs with around 140 intervention schools. Large-enough trials will of course be more expensive, reducing by around 75% the number of EEF trials. This sounds costly, but it must be kept in mind that 93% of such trials have estimated effect sizes below their MDES and/or 40% were uninformative. Conducting fewer, larger trials is therefore likely to increase, not decrease, the number of informative findings.

One potential objection to this line of reasoning is that effect sizes of 0.08 are not worth detecting, because they imply a poor cost-benefit ratio and would therefore provide no useful information anyway. For some interventions, this will be true. For others however, such as text messaging interventions, costs can be as low as 1$ per pupil, implying that even interventions with small effects would be economically worthwhile implementing (Kraft, 2018). That larger RCTs can detect quite

small effects therefore only strengthens the argument that they are a feasible and affordable way of making trials more informative, across a wide range of interventions.

NOTE

[1] Based on random effects meta-analysis using one impact estimate per trial. Where a trial reported a maths impact estimate I use this, as it is likely subject to less measurement error (Rhead, Black, & Pinot De Moira, 2018). Where no maths impact estimate was available, I used a random non-maths impact estimate.

REFERENCES

EEF (2019, March 19). How do we make EEF trials as informative as possible? [Blog post]. Retrieved from: https://educationendowmentfoundation.org.uk/news/eef-blog-how-do-we-make-eef-trials-as-informative-as-possible/

Kraft, M. (2018). Interpreting Effect Sizes of Education Interventions. *Brown University Working Paper*.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: should we be concerned? *Educational Researcher*.

Rhead, S., Black, B., & Pinot de Moira, A. (2018). *Marking consistency metrics: an update*. Coventry: Ofqual.