The trials of evidence-based education: the promises, opportunities and problems of trials in education, by Stephen Gorard, Beng Huat See and Nadia Siddiqui, Abingdon, Routledge, 2017, 200 pp., £24.99 (paperback), ISBN 9781138209664

In the last six years, UK education research funding has been transformed. The Education Endowment Foundation (EEF), established by the Sutton Trust in 2011 with £125 million of funding, has funded more than 145 randomised controlled trials (RCTs), each costing an average of around £500k (EEF, 2015). By way of comparison, since 2012 the Economic and Social Research Council has spent just £4.1 million on open-call education research grants (ESRC, 2018). This makes the EEF by far the largest UK funder of education research, and its focus on RCTs represents a major change for education research and for mathematics education in particular: of the 263 articles published in the eight leading mathematics education journals in 2012, only eight reported studies with random allocation into groups (Alcock, Gilmore, & Inglis, 2013). Such a major change calls for reflection. Is this new focus on RCTs a positive development?

Gorard, See, and Siddiqui's (2017) book offers just such reflection. Its authors strongly support the increased use of RCTs, arguing that "much of the published research on education is of such poor quality that it might do more harm than good" (p. 4), that mostly it "is of no consequence or use for any real life purpose" and therefore that it "can safely be ignored" (p. 5). They argue that most studies do not involve randomisation into groups (so causality cannot be inferred), that too many educational interventions are evaluated by reference to anecdotes or student satisfaction surveys, and that many educational research designs lack appropriate comparison or control groups. Thus, they claim, more well-conducted RCTs are needed if we are to draw genuinely causal conclusions about "what works".

Given these comments, one might expect Gorard et al. to be delighted by the recent emergence of the EEF. But the authors' position is more nuanced. While they believe that the increase in RCTs has led to "considerable progress" (p. 18), they spend much of the book critiquing the methods commonly adopted by the EEF's RCT researchers and advocating their own alternatives. These alternative methods will, Gorard et al. suggest, allow education researchers to establish what works. I agree with Gorard et al.'s overall view about the need for better evidence, but see serious problems with both their critique of traditional RCT methods and their apparent philosophical views on the purpose of research. These problems are explored in the remainder of this review.

The RCT method

Gorard et al.'s criticisms of existing approaches to RCTs come in a chapter entitled "problems, abuses and limitations in the conduct of trials". This chapter commences with observations about how bodies such as the EEF allocate funding, before arguing that missing data is a larger problem in educational RCTs than is suggested in EEF guidelines. But the chapter really gets going when it considers null hypothesis significance tests (NHSTs), the primary quantitative method used by behavioural scientists when analysing data from RCTs (as well as other research designs).

Gorard et al. make two main criticisms of NHSTs. First, they assert that a fundamental assumption of all NHSTs is that study participants are randomly sampled from the population of interest. Since this is never the case in real-world research (participants who refuse consent are never included, for instance), they conclude that "significance tests should never be used",

and that any p value calculated from real data "does not and cannot mean anything" (p. 22). This is a big claim: if the argument were even partially correct, then large sections of the education and psychology literature would be meaningless.

Gorard et al.'s second criticism is that researchers who use such tests fail to understand what a p value is. They state that p values quantify how likely the observed data (D) is given that the null hypothesis (H_0) is true; in symbols, $P(D|H_0)$. But, they say, what researchers "actually want is the probability of the nil-null hypothesis being true given the data observed". Since $P(D|H_0) \neq P(H_0|D)$, they conclude that NHSTs "just do not work" (p. 28), and that their use should be abandoned. In an extraordinary passage that should have no place in serious academic writing, the authors assert that "significance testing appears to derive from a psychological flaw among its advocates and defenders" (p. 28). Strong words! Unfortunately, Gorard et al.'s analysis is extremely simplistic. While there are many well-formulated critiques of NHSTs, and while alternative approaches could be more appropriate in some circumstances,[1] neither of Gorard et al.'s criticisms, at least in the form presented here, are persuasive.

With respect to the first criticism, Gorard et al. subtly mischaracterise p values. A p value quantifies the chances of observing the test statistic (or one more extreme) if *all* model assumptions held. One assumption certainly is that the null hypothesis is true. But there are others. Some concern how the data were collected (i.e. that there was no optional stopping based on data peeking). Others – if a classical test is used – may concern how the sample was selected. A small p value indicates a small probability that the test statistic would be as extreme as the observed value if *every one* of these assumptions were correct. As Greenland et al. (2016) point out, a small p value cannot tell us which (if any) of the model assumptions are incorrect. For that we need judgement. This means that for Gorard et al.'s critique to be as powerful as they suppose, they need to demonstrate that non-random sampling, coupled with a classical significance test, is likely to lead to artificially small p values.

This question can be tackled by simulating non-random samples and comparing the p values obtained from classical significance tests with those from randomisation tests. This is because, contrary to the claim made by Gorard et al., it is not true that *all* forms of NHST assume random sampling: randomisation tests explicitly avoid this assumption (e.g. Edgington & Onghena, 2007; Todman & Dugard, 2001). Conducting such a simulation is easy to do yourself. (See https://doi.org/10.6084/m9.figshare.6016247.v1 for the code from a simple simulation I conducted. It demonstrates that, in at least some situations, violating the assumption of random sampling is not a serious issue.) However, there is also a large literature on this topic. For instance, Edgington (1966) argued that the best way of thinking about classical tests is that they approximate randomisation tests, stating that "the closeness of the approximation under certain conditions has been shown theoretically (Silvey, 1954; Wald & Wolfowitz, 1944) and by numerical examples (Eden & Yates, 1933; Fisher, 1935, Section 21; Kempthorne, 1952, p. 152; Pitman, 1937; Welch, 1937)" (p. 487). Even if Gorard et al. were unpersuaded by Edgington's claim that classical tests approximate randomisation tests, they could simply dispense with p values calculated from classical tests entirely, and replace them with those derived from randomisation tests.

In sum, Gorard et al.'s first objection to NHSTing is unconvincing. But what of the second? They argue that researchers want to know $P(H_0|D)$, when NHSTs provides only $P(D|H_0)$. But this argument is a form of Lakens's (2017) "statistician's fallacy": Gorard et al. claim to know "what analysts want" (p. 25), but this assumes both that all analysts want the same thing in all situations, and that Gorard et al. can intuit what this is without empirical investigation. Speaking for myself, they are simply wrong. When I calculate a p value I do not want to know $P(H_0|$

[1]For a balanced account of the three main approaches to statistical inference, I recommend Dienes's (2008) excellent book.

D). I'm sure about this because p values are a frequentist concept, and within the frequentist approach $P(H_0|D)$ does not exist.

What do I mean? Frequentists define probabilities in terms of long-term frequencies. We say that the probability of a fair coin landing tails is 0.5 because, in the long run, the frequency of tails will be half the total number of coin tosses. In symbols, the probability is $P(T) = \lim_{n\to\infty} (n_T/n)$, where n_T is the number of tails observed and n is the total number of coin tosses. But this definition does not apply in the case of H_0, which either is a property of the mechanism that generated the data, or is not. There is no "long run" over which to observe the frequency of H_0, so it is meaningless to talk about its frequentist probability. Thus, a researcher who wants to know $P(H_0|D)$ should not be calculating a p value. But this is not for the reason Gorard et al. give, it is for the more fundamental reason that the desired quantity does not exist within the paradigm in which p values sit.

So what *do* I want to know when calculating a p value? Like all good Neyman-Pearson hypothesis testers, I want to control my long-term error rates by adopting a fixed decision rule. I know that if I follow the rule "reject the null hypothesis whenever $p < .05$", then *in the long run* I won't reject true nulls more than 5% of the time. This tells me nothing about any particular null, but it does help me (or a funder) have confidence in my (or their) whole body of work. The problem with Gorard et al.'s critique of NHSTing is not that it's wrong: it's a fair, albeit incomplete, critique of an incoherent version of NHSTing. The problem is that it doesn't address how NHSTs works if conducted in line with the Neyman-Pearson paradigm.

But what of researchers who, like Gorard et al., do want to know the "probability" of a hypothesis such as H_0? Clearly NHSTs aren't for them, but what should they do instead? Gorard et al. offer a selection of recommendations in their fourth chapter. They suggest that sample sizes should no longer be based on power analyses, that there should be greater use of standardised effect sizes (but see Simpson, 2017), and that researchers should report a figure they call the "number needed to disturb" (NNTD). I found some of the arguments offered in support of these suggestions to be genuinely strange. For instance, Gorard et al. argue that power calculations (fundamental to the Neyman-Pearson approach to inference) are "internally contradictory" (p. 40). Their reason boils down to the claim that it is contradictory to calculate both $P(A|B)$ and $P(A|C)$ unless B and C are identical, which is palpably absurd.[2] An equally questionable claim is the assertion that "effect sizes are theoretically independent of sample sizes" (p. 32), which would be true only if researchers did not choose their sample size based on the expected effect size (Simonsohn, 2017).

Perhaps the oddest recommendation is that researchers should focus on the NNTD rather than reporting inferential statistics such as p values. The idea is that the NNTD quantifies "the number of counterfactual cases needed to disturb the finding" (p. 45), where counterfactual cases are defined to be new participants whose scores are one standard deviation away from the smaller group's mean. Quite apart from the arbitrariness of this definition (why one standard deviation?), the NNTD is vulnerable to every criticism that Gorard et al. level vociferously at NHSTs. In particular, a NNTD quantifies $P(H_0|D)$ no more than a p value does. Since $P(H_0|D)$ is apparently what Gorard et al. want to know, this seems a fatal weakness. I wondered why Gorard et al. did not simply suggest that researchers adopt an approach to statistical inference that *does* allow the probability of a hypothesis to be calculated (albeit defined in a non-

[2]Gorard et al. write: "a power calculation starts by envisaging a non-zero effect size (the estimated effect of the treatment in an RCT). The researcher assumes this non-zero effect size as the bedrock for the calculations that follow. The calculations themselves are also predicated on a significant test, which was shown in Chapter 3 to assume as the basis for its own calculation that there is no effect sizes (the nil-null hypothesis). Put another way, the p-values generated by significance tests assume an ES of precisely zero. Both of these initial assumptions cannot be true in the same calculation, by definition. Therefore, 'power' does not make sense" (p. 40).

frequentist paradigm where probabilities are subjective degrees of belief). Researchers who really want something like $P(H_0|D)$ from their statistical analyses should adopt a Bayesian approach to probability and report Bayes factors or Bayesian credible intervals, not NNTDs.

The RCT philosophy

The second half of Gorard et al.'s book reports a series of trials conducted by the authors, and is designed to illustrate how their methods might work. Some of the trials focus on the transition to secondary school, some on early literacy and numeracy, some on teaching philosophy to children, and so on. The disparate nature of these topics stands out. Gorard et al. conduct trials across all educational domains and, from the evidence of the reports in these chapters, see little purpose in getting to grips with the relevant literatures. In fact, they are explicit about this, explaining that they have no interest in educational theory: "theoretical explanations appear satisfying but are unnecessary when assessing 'what works'" (p. 101).

Can this be right? Are theoretical explanations unnecessary for educational research? This raises the question of what research is for. Is it, as Gorard et al. assert, simply to catalogue facts about what works? Or is it to develop our understanding so that we can *predict* what will work? These two approaches to research – fact-accumulation versus theory-building – were articulated by Mook (1983) in his classic defence of externally invalid psychology experiments. Mook argued that research is often designed not to establish facts about what works here and now, but to test and refine theories which can later be applied to numerous real-world settings.

Some go further than Mook, arguing that untheorised RCTs like those reported by Gorard et al. do not even allow us to establish "what works". For instance, Cartwright (2011) accepted that RCTs are ideal for establishing claims of the form "the treatment caused the outcome in some members of the study", which she referred to as the "it-works-somewhere" claim. But, suggested Cartwright, in almost all real-world situations this is not the claim we're interested in. Instead we want to know if it will work for *us*. Cartwright and Hardie (2012, p. 80) gave a simple example of how a lack of theoretical understanding can lead to mistakes. They explained how a programme shown to improve pregnant women's nutrition in Tamil Nadu was transferred to Bangladesh. A major component of the intervention was nutritional counselling for the pregnant women. The programme was effective in Tamil Nadu, but a new RCT in Bangladesh found no effect. Why? It seems that in Tamil Nadu women are traditionally responsible for food shopping, whereas in rural Bangladesh it is typically men who go to food markets. The programme could never have been successful, as the mechanism upon which it relied (pregnant women choosing healthier food when shopping) did not exist in the new context.

How can we move from an it-works-somewhere claim to an it-will-work-for-us claim? We need theoretical understanding of the causal mechanism that links intervention and outcome. This is what allows us to assess whether the necessary factors are present in our situation, and to anticipate whether malign factors might disrupt that mechanism. In other words, we need exactly the kind of theoretical understanding that Gorard et al. suggest is unnecessary. Indeed, when viewed through this "what is the mechanism?" lens, some conclusions drawn by Gorard et al. look like missed opportunities. For example, when summarising the results of six literacy trials, Gorard et al. write "It is clear that simply using commercial software to teach literacy does not work, and this should be avoided" (p. 101). But it is clearly unjustified to say that teachers should avoid using all commercial software to teach literacy: the apparent failure of the packages tested was surely related to their pedagogical approach rather than to their

commercial origins. Without understanding the hypothesised theoretical mechanism by which the software was designed to improve literacy, it is hard to draw useful conclusions from such trials.

Conclusion

In all, Gorard et al.'s book is interesting but flawed. Its authors are right to draw attention to the new reality of education research and its focus on RCTs; and they are right to critically evaluate the progress made via this new approach. But many of their criticisms are simply unconvincing. Gorard et al. advance the radical claim that the majority of education and psychology research is both statistically and philosophically misguided but, surprisingly and disappointingly, they fail to anticipate and discuss some fairly obvious objections to their arguments.

Notes on contributor

Matthew Inglis is a Reader in Mathematical Cognition in the Mathematics Education Centre at Loughborough University. He is interested in understanding the processes involved in numerical and mathematical thinking, and how these can be promoted through education.

References

Alcock, L., Gilmore, C., & Inglis, M. (2013). Guest editorial: Experimental methods in mathematics education. *Research in Mathematics Education*, *15*, 97–99.

Cartwright, N. (2011). A philosopher's view of the long road from RCTs to effectiveness. *The Lancet*, *377*(9775), 1400–1401.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford: Oxford University Press.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Basingstoke: Palgrave Macmillan.

Eden, T., & Yates, F. (1933). On the validity of Fisher's *z*-test when applied to an actual sample of non-normal data. *Journal of Agricultural Science*, *23*, 6–16.

Edgington, E., & Onghena, P. (2007). *Randomization tests*. London: CRC Press.

Edgington, E. S. (1966). Statistical inference and nonrandom samples. *Psychological Bulletin*, *66*, 485–487.

Education Endowment Foundation (EEF). (2015). *Annual Report 2014/15*. Education Endowment Foundation.

ESRC. (2018). *ESRC Application and Success Rate Data*. Retrieved from http://www.esrc.ac.uk/files/about-us/performance-information/application-and-success-rate-data/

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver & Boyd.

Gorard, S., See, B. H., & Siddiqui, N. (2017). *The trials of evidence-based education: The promises, opportunities and problems of trials in education*. Abingdon: Routledge.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*, 337–350.

Kempthorne, O. (1952). *The design and analysis of experiments*. New York: Wiley.

Lakens, D. (2017). The Statisticians' Fallacy. *The 20% Statistician Blog*. Retrieved from http://daniellakens.blogspot.co.uk/2017/11/the-statisticians-fallacy.html

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*, 379–387.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations: III. The analysis of variance test. *Biometrika*, *29*, 322–335.

Silvey, S. D. (1954). The asymptotic distributions of statistics arising in certain nonparametric tests. *Proceedings of the Glasgow Mathematics Association*, *2*, 47–51.

Simonsohn, U. (2017). The Funnel Plot is Invalid Because of This Crazy Assumption: r(n,d)=0. Data Colada: Thinking about evidence and vice versa, 58. Retrieved from http://datacolada.org/58

Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised effect sizes. *Journal of Education Policy*, *32*, 450–466.

Todman, J. B., & Dugard, P. (2001). *Single-case and small-n experimental designs: A practical guide to randomization tests*. New York, NY: Psychology Press.
Wald, A., & Wolfowitz, J. (1944). Statistical tests based on permutations of the observations. *Annals of Mathematical Statistics*, *15*, 358–372.
Welch, B. L. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika*, *29*, 21–52.

Matthew Inglis
Mathematics Education Centre, Loughborough University
m.j.inglis@lboro.ac.uk

https://doi.org/10.1080/14794802.2018.1481451