Routledge
Taylor & Francis Group

Critical Studies in Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rcse20

Challenging the gold standard consensus: Randomised controlled trials (RCTs) and their pitfalls in evidence-based education

Juan David Parra & D. Brent Edwards Jr

To cite this article: Juan David Parra & D. Brent Edwards Jr (18 Feb 2024): Challenging the gold standard consensus: Randomised controlled trials (RCTs) and their pitfalls in evidence-based education, Critical Studies in Education, DOI: 10.1080/17508487.2024.2314118

To link to this article: https://doi.org/10.1080/17508487.2024.2314118

CRITICAL STUDIES IN EDUCATION
https://doi.org/10.1080/17508487.2024.2314118

Routledge
Taylor & Francis Group

Challenging the gold standard consensus: Randomised controlled trials (RCTs) and their pitfalls in evidence-based education

Juan David Parra [a] and D. Brent Edwards Jr [b]

[a]Assistant professor Instituto de Estudios en Educación (IESE), Universidad del Norte, Barranquilla, Colombia; [b]Department of Educational Foundations, University of Hawaii at Manoa, Honolulu, HI USA

ABSTRACT

This paper seeks to raise awareness among educational researchers and practitioners of some significant weaknesses and internal contradictions of randomised control trials (RCTs). Although critiques throughout the years from education scholars have pointed to the detrimental effects of this experimental approach on education practice and values, RCTs are considered the *gold standard* for assessing the impact of education policies and interventions. By drawing on the approach of immanent critique, we elucidate substantial argumentative gaps between the assumptions and applications – that is, between the theory and reality – of RCTs in empirical research. This kind of analytic exercise complements existing critiques from outside the experimental discourse based on moral and epistemic principles. The present paper, in contrast, contributes to the literature by highlighting internal limitations and contradictions that can be seen by probing the logic espoused by those who are proponents of RCTs. In fleshing out our argument, we seek to encourage more informed and critical engagement by educators, policymakers, and researchers, among other stakeholders, when they are confronted with proposals for education programmes and reforms supported by findings from RCTs.

ARTICLE HISTORY
Received 17 January 2023
Accepted 31 January 2024

KEYWORDS
Evidence-based education; immanent critique; RCTs; impact evaluation; evidence-based policy

Introduction: the pitfalls of experiments in evidence-based education

This paper aims to assist education scholars, practitioners, and policymakers in becoming more critical consumers of research and claims based on randomised controlled trials, or RCTs. We see this goal as particularly important in light of the exponential growth and high-profile endorsement of RCTs to inform evidence-based education in the last two decades (Barrenechea et al., 2022, Connolly et al., 2018 Simpson, 2019,). Given their background and inspiration in clinical experimental research (Bhatt, 2010), exponents often portray RCTs as a *gold standard*, that is, a method that 'always provide[s] the strongest evidence for causality and for effectiveness' (Deaton & Cartwright, 2018). Consequently, RCTs are placed at the top of the hierarchy of scientific knowledge for improving the quality of education in different settings (Banerjee & Duflo, 2009, Slavin,

CONTACT Juan David Parra jparrad@uninorte.edu.co Assistant professor Instituto de Estudios en Educación (IESE), Universidad del Norte, Barranquilla, Colombia

2002, Walters, 2009). This perspective holds that evidence generated by RCTs should inform efficient planning by allowing to focus educational investments on *what works* (Gertler et al., 2016).

Such is the case, despite significant questioning from educators on whether this type of experimental inquiry responds to the moral and epistemic principles of education (Biesta, 2007, McKnight & Morgan, 2019). For instance, in his influential critique of the *what works* paradigm, Biesta (2007) contends that the RCT approach 'is at least insufficient and probably misplaced in the case of education, because judgment in education is not simply about what is possible (a factual judgment) but about what is educationally desirable (a value judgment)' (p. 10). Some commentators have also delved into the ontological premises of RCTs (e.g. by pointing out its ideological biases and commitments to positivism), resulting in calls for strengthening the influence of alternative approaches for informing evidence-based education (Barrenechea et al., 2022, Gale, 2017, Joyce & Cartwright, 2020, Tikly, 2015, Wrigley & McCusker, 2019). Such critiques find at least some validation in recent quantitative meta-analyses which suggest that, given statistical noise – i.e. considerable variability in experimental estimations of the impact of similar educational interventions across contexts – RCT studies are actually a poor basis for informing decision-making (Evans & Popova, 2016, Lortie-Forgues & Inglis, 2019, Simpson, 2017, Vivalt, 2020).

This paper contributes to the critique of RCTs. However, our contribution consists in taking an alternative argumentative path from the ones outlined above. Specifically, we go beyond critiques from education scholars that question the relevance and appropriateness of RCTs in educational research. While necessary, these critiques appear not to have had their desired effect given that RCTs continue to be very influential in education reform. Echoing some recent sociological literature (Berman, 2022, Gale, 2017, Pearce & Raman, 2014), we hold that at least part of the explanation for this circumstance lies in the fit of the underlying experimental paradigm with the common sense of the contemporary technocratic-oriented policymaker. Under such situation, it seems vain to challenge this approach using exclusively arguments emerging from alternative paradigms or normative principles – like Biesta's (2007) questioning of experimental methods for potentially deterring democratic deliberations around education policy – To this end, our analysis finds inspiration in the critical realist notion of an *Achilles Heel immanent critique*, which, in words of one leading exponent of this philosophy, 'specifies that criticism of an idea or a system should be internal, that is, involve something intrinsic to what (. . .) is being criticised' (Bhaskar, 2016, pp. 1–2). Critique of this nature might be conducive to illuminating contradictions or inconsistencies within a common sense theoretical position (Isaksen, 2018, O'Mahoney & Vicent, 2014), improving, therefore, the odds of persuading enthusiastic practitioners about why they should become more critical consumers of experimental research in education.

We can hence think of an immanent critique 'as a method of argumentation' (Isaksen, 2018, p. 98) which begins by considering 'minor premises accepted or implied by the account that it seeks to situate, correct or refute (. . .) to demonstrate that the account is internally inconsistent or beset with problems that cannot be solved in its own terms' (Hartwig, 2008, p. xiv). Following this methodological orientation, our critical overview of the strengths and limitations of RCTs in educational research does not begin by citing scholars that criticise them based on epistemic considerations external to the

experimental approach. Instead, we focus on examining the internal consistency of rhetoric behind the supposed merits of randomisation as portrayed by leading RCT exponents (e.g. as a statistical technique to safeguard objectivity in evidence-based education discussions). As such, our purpose is not to contradict previous critiques of RCTs emerging from education scholars but instead to contribute to their effort by taking a different tack, one that assesses the merits of RCTs from within.

The structure of the paper reflects the spirit of the analysis mentioned just above by first making a detailed presentation of the basic tenets and assumptions of experimental evaluations (such as RCTs) regarding their ability (through randomisation of individuals into treatment and control groups) to ensure *clean* (or unbiased) measurements of the impacts of (educational) interventions. This section seeks to shed light on the approach's basic premises – defined in its *own terms* – in order to ensure that readers have the necessary background information to follow the arguments that are presented subsequently. After clarifying the basic foundations behind the ideas and merits of randomisation, we then explore and judge the internal consistency of two (types of) standards that experimental researchers conceive as 'obvious requisites for the worth of an RCT' (Zhang et al., 2015, p. 32). We refer here to i) *internal validity*, or the possibility of estimating the impacts of education interventions without statistical bias, and ii) *external validity*, or the ability to extrapolate from these estimations to predict the effect of interventions in multiple contexts.

The discussion in these sections focuses on developing four principal arguments, all pointing to the impossibility of purely statistical solutions (a central premise of the randomisation rhetoric) to the underlying problems of this method. While these arguments are relevant to RCTs applied in any field, we focus in our paper on educational research. Although their exponents argue the opposite, we contend that: i) RCTs cannot mitigate threats posed by attrition (which happens when comparison groups created through randomisation lose comparability over time); ii) RCT reliance on significance testing (and avoidance of social theory) to validate their findings is misplaced; iii) the (inevitable) unbalancing (i.e. incomparability) of treatment and control groups means that RCT results are not generalisable; and iv) given that RCTs, by themselves, add nothing to the identification of causal mechanisms triggered by (educational) interventions, their results are uninformative for decision making (both within and beyond a given context).

Finally, and since our intended audience includes policymakers and education practitioners, we close our discussion by highlighting some real-world risks behind continuing the uncritical endorsement of RCTs to inform education policy. Here, we engage some examples from recent literature to argue that narrow technocratic preferences toward experimental research might contribute to reproducing the vested interests of organisations that sponsor specific knowledge production and mobilisation, with potentially adverse effects on the opportunities for transformative education system reform, particularly in low and middle-income contexts.

Getting the impact right: the foundations of RCTs for impact evaluation

In evidence-based education discussions, RCTs attempt to answer the question: '[w]hat is the impact (or causal effect) of a programme on an outcome of interest' (Gertler et al.,

2016, p. 8). The methods used are known as experimental impact evaluations and they are based on the creation of a counterfactual (Banerjee & Duflo, 2009). The fundamental idea is to apply statistical principles to determine 'what the outcome would have been for programme participants *if they had not participated* in the programme' (Gertler et al., 2016, p. 6, *emphasis added*). In other words, experiments seek to replicate a scenario in which one can explain what would have happened with an individual, vis-a-vis a specific indicator or outcome, if s/he had not received the treatment.

The argument for the necessity of using experimental (i.e. randomised) research to identify the impacts of specific social interventions correctly usually starts by pointing out that simply observing how indicators of (non-randomised) sets of participants evolve over time is insufficient to answer causal questions statistically. Consider the hypothetical case of an evaluation of a teacher-training programme that introduces new teaching tools to enhance student achievement. Identifying positive associations between teacher participation in training sessions and changing student achievement outcomes in some schools does not mean the former (i.e. the training) is responsible for the observed changes. There are many reasons why students might score better on exams or standardised tests that have nothing to do with teacher training, like family socioeconomic status, race, gender, and peer group effects. The observed changes in outcomes of interest may therefore be explained more due to those pre-existing differences between students' backgrounds than to the impact of the intervention itself (Banerjee & Duflo, 2009).

An alternative for assessing the causal impact of the hypothetical teacher-training programme is to conduct an RCT. Following the spirit of clinical trials in medical research, such a strategy entails building comparison groups consisting of the intervention's beneficiaries (treatment group) and non-beneficiaries that have similar background characteristics to the participants of the intervention (control group). The challenge here is to ensure both groups are statistically comparable so evaluators can confidently attribute any difference in pre and post-intervention measures of a variable of interest to the programme itself (Sedgwick, 2015). If, for instance, some teachers in either of the comparison groups belong to schools that benefit from other interventions that could affect students' engagement and learning (e.g. free school meals), the statistical balance between groups is compromised. The assessment of the interventions would thus suffer from what impact evaluators call *selection bias*, which happens 'when unobserved reasons for programme participation are correlated with outcomes' (Gertler et al., 2016, p. 33).

To address such statistical challenges, RCT advocates suggest selecting comparison groups before beginning the intervention by picking them randomly from a pool of potential candidates. This approach is based on a statistical rule called the *law of large numbers*, which establishes that, under random assignment, 'a sample average will approximate the average of the population from which it is drawn as the sample size grows larger' (Thomas & Chindarkar, 2019, p. 32). By endorsing this principle, it is argued that, 'in theory, the control groups generated through random assignment serve as *a perfect counterfactual* [of the treatment group], free from the troublesome selection bias issues that exist in all evaluations' (Baker, 2000, p. 2. *Emphasis added)*. This conclusion derives from the testable fact that if one runs a lottery to select (large enough[1]) subsamples (or groups of participants) of a larger population, all groups will have, on average, identical statistical background characteristics compared to the original population (Fashing & Goertzel, 1981, Hubbard et al., 2019).

Random allocation is hence at the core of the justification for why RCTs are supposedly better equipped than alternative approaches in establishing clean impact estimations. In our hypothetical example of a teacher training programme, endorsing randomisation for its evaluation entails using a lottery to assign (randomly) teachers to treatment and control groups before the intervention, with these two groups drawn, for example, from a pool of schools with historically low achievement and with similar probabilities of exposure to other policy interventions. For proponents of RCTs, clean impact estimates will be obtained because 'by construction the two groups were identical at the baseline' and, therefore, they will be 'exposed to the same external environmental factors over time' (Gertler et al., 2016, p. 69). The theory says that selection bias in the creation of comparison groups will be avoided and, consequently, 'differences between the treatment groups in outcomes at the end of the trial will be due to differences in treatment and not to differences in baseline characteristics, thereby permitting the inference of causality to be ascribed to a treatment' (Sedgwick, 2015, p. 1).

We have now delineated the basic methodological foundation of experimental evaluations and their alleged contributions to evidence-based education. However, before proceeding, something worth highlighting here – because of its relevance to the argument in the next section – is that the rationale behind the benefits of baseline randomisation in RCTs is purely statistical, as opposed to being informed by social theory (Hubbard et al., 2019). As Surendran and Kumar (2020) explain, RCTs derive from an empiricist philosophical position according to which 'reality is circumscribed by observable means alone'(p. 1), which sustains the belief that researchers do not require theory to specify (or define what information to add in) their statistical models. As such, RCT users argue that they can test existing theories based on their results without needing to draw on those theories to guide data collection (Kvangraven, 2020). This supposedly theory-free nature of the method is a feature that its foremost exponents celebrate because they argue that it shields it from the subjective preferences of researchers, hence making it scientifically robust (Banerjee & Duflo, 2009).[2] However, as will be seen, this aversion to theory is actually a weakness of this approach to programme and policy evaluation.

The challenge of internal validity in RCTs

In (quantitative) research and evaluation, internal validity refers to 'the question of whether a given [research] design successfully uncovers causal effects for the [sampled] population being studied' (Angrist & Pischke, 2009, p. 151). In line with the methodological discussion above, randomisation to create comparison groups in RCTs is ultimately about safeguarding internal validity as it intends to remove statistical noise (e.g. the influence of background variables) from impact estimates. Or, as Gertler et al. (2016) put it, '[the] degree of comparability between treatment and comparison groups is central to the evaluation's internal validity and is therefore fundamental to assessing a programme's causal impact' (p. 22).

In this section, we present two critiques related to the ability of RCTs to safeguard internal validity *by themselves* (i.e. through randomisation). Specifically, we will discuss two inconsistencies or contradictions internal to the experimental evaluation rhetoric

and practice, particularly in educational research, that prevent the materialisation of the advantages of baseline randomisation from identifying clean impact estimators in assessment projects. These are i) the problem of attrition and ii) the problem of *faux exogeneity*.

The practical difficulty (impossibility) of solving attrition problems

Attrition occurs when 'units assigned to the treatment or control group drop out of the experimental study' (Thomas & Chindarkar, 2019, p. 37) thus creating technical difficulties in avoiding estimation biases due to a lack of balance between groups. Attrition in educational interventions might respond to various circumstances, such as participants (e.g. in our hypothetical example) moving to other schools (Henneberger et al., 2023). This is a significant problem with RCTs because, as Deaton and Cartwright (2018) point out, attrition is something that 'almost always occurs in practice' (p. 17). For instance, Henneberger et al'.s (2023) has found, using State Longitudinal Data from the United States, 'that researchers using K-12 school-based samples should plan for attrition rates as high as 27% during middle school and 54% during elementary school" (p. 1).

One common suggestion by experts to address sample loss due to attrition in RCTs is to 'enrol more participants than the minimum required (. . .) to compensate for expected withdrawals' (Negida, 2017, no page number; see also Anderson et al., 2021). Intuitively, replacing individuals lost with similar individuals available due to oversampling is a reasonable strategy to re-balance both comparison groups after initial randomisation. The argument states that matching exercises based on the observed background attributes of the remaining individuals can assist evaluators in verifying whether the baseline characteristics of those who dropped out of the sample are statistically equal to those that remained (Austin & Stuart, 2015, Gertler et al., 2016). In matching exercises, the idea is that one 'statistically reweights the data (. . .) to correct for the fact that a portion of the original respondents is missing' (Gertler et al., 2016, p. 71).

But, to what extent is standard practice related to re-balancing in RCT studies compatible with the premises of randomisation? We answer that it is largely incompatible. According to the statistical theory outlined in the previous section, pre-treatment randomisation ensures that individuals in both the treatment and control groups have statistically identical observable and non-observable background characteristics without the need to endorse any particular social theory for guiding this process (e.g. a theory that indicates essential individual attributes to take into consideration when recruiting participants in the first place). It is therefore evident that using matching strategies to re-balance groups post-randomisation *based only on the observable characteristics* (and, more specifically, only those characteristics that are available as variables to be used in the matching exercise) contradicts the supposed benefit of randomised sampling. The point is that nothing assures that the replaced individuals in the new sample will remain *entirely* balanced, hence generating post-randomisation selection bias (Deaton & Cartwright, 2018; Hernán et al., 2013).[3]

Still, for the argument's sake, let us assume that evaluators resort to matching methods and regain balance after attrition. That means that by using additional statistical techniques, they manage to have, once again, comparison groups to measure the impact of an intervention eliminating selection bias. Nevertheless, precisely because that re-balancing effort ought to use, by definition, only observable variables, another problem that arises

here is that nothing ensures that the resulting groups are still representative *of the original population* (in terms of unobservables) targeted by the specific intervention under assessment (Barry, 2005). As Greenberg and Barnow (2014) put it, '[the] key problem resulting from sample attrition is that it is unlikely to be random. This can make the remaining sample unrepresentative of the group originally included in the research sample'. (p. 368). Stated plainly, attempting to re-balance groups due to selective attrition (for unknown reasons) means that the average background characteristics of the new (or re-calibrated) comparison groups may now differ from those of the original population targeted by the assessed intervention.

The significance of this limitation is clear if one recalls that RCTs are often performed to understand the impact of an intervention on specific populations (e.g. underperforming students in a whole school district). However, in attempting to safeguard internal validity, RCT researchers jeopardize their ability to present their findings as the impact on (or with relevance for) that specific population targeted by an intervention (which was the promise of using an experimental technique in the first place). Instead, they can only assert that their findings apply to specific subsets of individuals within their sample – those who, for non-random reasons, remained.

Statistical significance testing and the problem of faux exogeneity

Quantitative researchers (and impact evaluators) perform statistical tests to assess the likelihood that their results are due to chance. One such test is known as the *t-test*, through which the t-statistic is computed. A typical assertion made about RCT-based impact evaluations states that:

> . . . by the value of the t-statistic (. . .), the difference between [a given value across subgroups] before and after the programme is *statistically significant*. This means that you find strong evidence against the claim that the true difference [in the given outcome of interest] (. . .) before and after the intervention is zero. (Gertler et al., 2016, p. 56)

The confirmation of this difference in an outcome indicator between treatment and control groups after an intervention depends on the value of the t-statistic, which is usually interpreted by experimental evaluators using another statistical parameter known as the *p-value*. Frequently, researchers use the threshold of $p < 0.05$ (or five per cent probability) to *refute the null (or default) hypothesis* that a difference in the magnitude of an indicator of interest between groups is zero. Suppose a *p-value* is over the five per cent threshold. In that case, it is held that it is unlikely that there are statistically significant differences in the mean outcome indicator after the intervention between the treatment and the control group. Crucially, in RCTs, testing for statistical significance is the principal means for determining if an intervention had causal effects.

Now, from a strictly technical viewpoint, Kim and Park (2019) remind us that '[t]he conditions required to conduct a t-test include (. . .) appropriate sample size, and normal distribution of data' (Kim & Park, 2019, p. 332). As noted, RCT proponents hold that, under ideal settings, one merit of randomisation is that it enables the identification of representative subsamples of a population, each characterized by average attributes that adhere to a normal distribution (Morgan & Rubin, 2012).[4] Assuming the balance (across treatment and control groups) is maintained over time, the resulting comparison groups

after the intervention will, in theory, continue to hold the properties of normal distributions and statistical representation of the original population. Consequently, conditions are seemingly in place for t-tests to confirm statistical significance in experimental research.

However, Carver (1978) wrote decades ago in the *Harvard Education Review* a decisively critical article making the case against statistical significance testing in educational research, which addressed this last point. Experimental evaluators, he argued, can never have certainty that both sampled groups continue to represent the same population after the assessed treatment occurs; yet, as he noted, statistical testing 'says that we will assume that they do' (Carver, 1978, p. 381). In his article, he illustrates his argument by reflecting on the testing of the impact of a school-based programme to improve reading skills in children by comparing test scores in literacy between randomly created samples X (the treated) and Y (the control group). Suppose the intervention indeed impacted the expected outcomes positively. His point is that in this case, 'sample Y still represents the original population, *but sample X no longer does*—it now represents another population called the experimental population' (p. 380, *emphasis added*). If the populations are now fundamentally different, what does the t-test (based on a comparison of values taken after the transformation) actually reveal?

Some more contemporary work on the subject matter by Barrett and Carter (2010) extends Carver's (1978) argument. For them, the problem has an ontological character rooted in the reductionist portrayal of humans – i.e. as passive, and non-reflective beings – exhibited by experimental evaluators:

> Human agency complicates matters enormously (. . .). It is often unclear what varies beyond the variable the researcher is intentionally randomising (. . .). As a result, impacts and behaviours elicited experimentally are commonly endogenous [i.e., specific] to environmental and structural conditions that vary in unknown ways within a necessarily highly-stylised experimental design. This faux exogeneity undermines the claims of clean identification due to randomisation. (Barrett & Carter, 2010, p. 524)

In this quote, the authors problematize the assumption that after a successful intervention, nothing of the treated population should change *except for* their intervention's outcome variable (Porter et al., 2017, Van Belle et al., 2016). In section one of this paper, we showed that making this argument is essential for the RCT methodological narrative because relaxing it – i.e. allowing for the possibility that other non-outcome variables change due to the intervention – contradicts the principle of post-treatment balance itself. After all, if something else aside from the outcome variable changed, that would indicate that groups after an intervention are not comparable anymore. Barrett and Carter's (2010) point is that RCT proponents assume a *false exogeneity* (i.e. they assume that only external factors explain people's behaviour within the intervention), thus creating the illusion of statistical control over estimation errors.

However, taking attributes of human agency (e.g. reflexivity, spontaneity) seriously in impact evaluations means acknowledging that, even if the original comparison groups examined in an RCT resulted from pre-treatment randomisation, there are unpredictable ways that contextual factors might influence participants' actions in (and during) an intervention (Barrett & Carter, 2020; Kaplan et al., 2020; Leamer, 2010). In our hypothetical teacher-training programme, imagine that, given their

experience and learning accumulated during the intervention, teachers in certain schools decided to create study groups to provide mutual support during the implementation process, enabling them to benefit from the intervention more than educators in other schools. This unexpected turn of events implies that the programme's effect might not follow a normal distribution after the treatment (Deaton & Cartwright, 2018, Hubbard et al., 2019). However, '[if] we wrongly assume that it does, we will make mistakes (. . .) by thinking that a large t-value indicates an effect of the treatment when, in fact, there is none' (Deaton, 2020, p. 5)[5]

Notably, in a more recent paper, Barrett and Carter (2020) regret that despite their warning a decade prior about the risks of *faux exogeneity*, they have not witnessed any significant evolution in how experimental evaluators conceive of human subjects. Indeed, contemporary social theorists have noted that denial of the role of human agency is typical of the crude empiricism represented in mainstream RCT practice and, moreover, that this crude empiricism threatens the relevance of the social (and particularly, educational) research generated to inform evidence-based education debates (Parra, 2018; Tikly, 2015).

The challenges of external validity

We now turn the discussion about RCTs to the quality standard of *external validity*, which refers to the possibility of 'the extrapolation of findings beyond the study sample to another population' (Williams, 2020, p. 63). According to mainstream evaluation handbooks, the use of random sampling in impact evaluations ensures 'that the evaluation sample accurately reflects the population of eligible units so that impacts identified (. . .) can be extrapolated to the population' (p. 73. *Original emphasis*). As reiterated throughout, this assumption is central to attempts to use impact evaluations to 'guide programme design and policy decisions' (Gertler et al., 2016, p. 63) beyond specific contexts where evaluations occur.

On this point, Williams (2020) invites us to think simultaneously about two intertwined concepts. The first concept is *generalisability*, which corresponds with 'whether evaluation results from a specific context will hold in unspecified other contexts (but without a specific destination context in mind)' (p. 11). The second concept is *applicability*, which concerns 'whether evaluation results from various other contexts will hold in the specific context in which a policymaker is working' (Ibid). Arguably, the second concept is of more interest to users of evaluations because it has implications for the decisions they make regarding (educational) interventions.

To illustrate RCT practitioners' expectations concerning these last two concepts, let us imagine that an evaluation team used an RCT to assess our hypothetical teacher-training programme and concluded it has causal effects on raising student achievement. Drawing that conclusion requires that the differences identified between comparison groups, both of which emerged in the first place as subsamples of the initially targeted population, are *generalisable* to the whole population. In addition to this assumption, the RCT research team could compare their results with RCTs of similar interventions in other contexts to present evidence of the potential future impact (Simpson, 2019). However, for Joyce and Cartwright (2020), these strategies are problematic when it comes to informing evidence-based education decisions. As they explain:

> [a]pparently, using RCTs justifies the inference from the internal validity in each study and consistent results across multiple studies (or faring well in a meta-analysis of study results) of this conclusion about what can be expected in general. Clearly this is a mistake (. . .) Adding up causal ascriptions, as the literature seems to recommend, amounts to induction by simple enumeration, which has long been condemned as a weak form of inference. (p. 155)

The problems with generalising experimental results connect to the debate about the merits of randomisation to address threats to internal validity (or estimation biases) in impact evaluations. In the previous section, we concluded that the lack of balance in post-treatment comparison groups undermining internal validity tends to be the rule instead of the exception in RCT studies. Therefore, comparing potentially flawed estimations to arrive at insights about the impact of a specific type of programme – or, in the quote above, *induction by simple enumeration* – will not correct for the problems within each study (Joyce & Cartwright, 2020, Simpson, 2017, Wrigley & McCusker, 2019). In the end, as Leamer (2010) insightfully asks: 'What's to extrapolate anyway? Our lack of knowledge?' (p. 35).

A second challenge to using insights from RCTs to inform policy decisions in a different context relates to the notion of *applicability*. For instance, policymakers in one setting might show interest in replicating a successful educational intervention conducted elsewhere. In Williams's (2020) view, this is a much less studied problem in the RCT literature vis-à-vis the challenge of generalisability [see also Stein et al., (2021)]. We hypothesise this is the case because this problem is intrinsic to experimental evaluations' epistemological grounds. Hence its solution is not merely statistical, as the following quote by Deaton (2010) helps to illustrate:

> RCTs of 'what works', even when done without error or contamination, are unlikely to be helpful for policy, or to move beyond the local, unless they tell us something about why the programme worked, something to which they are often neither targeted nor well-suited (. . .). For an RCT to produce 'useful knowledge' beyond its local context, it must illustrate some general tendency, some effect that is the result of [a] mechanism that is likely to apply more broadly (p. 448)

Echoing this message, Williams (2020) reminds us that any attempt to apply RCT results to inform policy in other contexts requires an 'understanding of the interactions between a policy's theory of change (. . .) and the actual characteristics of the context to which a policy is being transported' (p. 26). It is in that sense that Tikly (2015) cautions that 'RCTs may show at best surface level patterns of "what works" but they cannot tell researchers how or indeed why interventions "work" for different groups of learners'(p. 219). The crucial point to grasp is that one should not simply assume that the same programme design will yield the same effects elsewhere based on arguments that focus exclusively on the statistical aspects of evaluations, such as the potential for statistical representativeness of experimental results, as often suggested in classical impact evaluation manuals.

Therefore, if one accepts these arguments related to generalisability and applicability (and external validity in general), it becomes apparent that RCTs are not superior to alternatives in guiding decision-making. Surprisingly, even RCT proponents acknowledge that RCTs 'are often limited in providing insights into the channels by which the policy or programme affected the observed results' (Gertler et al., 2016, p. 13). This

evident contradiction internal to the discourse of influential proponents of the experimental approach to policy evaluation suggests, as we will further discuss in the last section of the article, tensions between the ideology and practice of mainstream educational researchers.

Why educational researchers and practitioners should remain critical of RCTs

So far, in line with a realist immanent critique, this paper has highlighted inconsistencies within the RCT rhetoric and practice. The internal contradictions we have highlighted as challenging the arguments behind the internal and external validity of RCTs might help explain why, as some authors suggest, the increase in their use across sectors has not signified an accumulation of knowledge to inform evidence-based education debates (Biesta, 2007, Joyce & Cartwright, 2020, Odom, 2021, Tikly, 2015). To the extent that Muller (2020) is right when he argues that 'there is no evidence that reliance on RCTs has produced better outcomes [in policy-oriented research] than alternative approaches' (p. 2), the gold-standard status enjoyed by experimental evaluations is indefensible.

Should we hence conclude that policymakers and education practitioners must abandon using RCTs? Answering that question requires a consideration of the benefits (which, according to the previous discussion, are unclear) and the costs of continuing to endorse the experimental paradigm, including their unintended consequences. Regarding the latter, various scholars have pointed out the power of the experimental movement worldwide to distort policy planning and prioritisation, particularly in low- and middle-income contexts (Khera, 2023, Muller, 2020), in the sense that 'emphasising randomised trials of micro-interventions could distract from basic structural and institutional changes needed for economic development and elimination of poverty' (Muller, 2020, p. 2). The concern raised by critics is that experimental evaluations are only feasible for narrow interventions. For example, Kumar (2016) discusses a case where the insistence of international organisations on hiring temporary teachers to improve educators' accountability in India responded fundamentally to the fact that, contrary to alternatives, this strategy was feasible for randomisation. In that sense, he raises the issue 'that preferring RCTs biases enquiries into, and efforts to reform, public systems' (p. 89).[6]

Also, beyond the significant financial expenses required to conduct RCTs (Khera, 2023), critics have raised concerns about considerable opportunity costs associated with their privileged status in programme and policy assessment. These costs result from how an overemphasis on experimental evaluations in evidence generation systematically undermines alternative research methods potentially better equipped to answer questions about causal mechanisms (or the channels by which interventions work to induce change). As indicated before, the problem is that this kind of knowledge is crucial in informing evidenced-based education debates (Barrenechea et al., 2022; Edwards, 2018; Joyce & Cartwright, 2018; Kvangraven, 2020; Maxwell, 2020; Tikly, 2015). Khera's (2023) recent comment speaks to this last point:

> Of course, validating an intuition or common sense requires research. But is an RCT the most cost effective method of doing so? There is pressure to use RCTs because they were

> projected as the gold standard (. . .) Privileging evidence from RCTs, over other evidence, raises questions because expensive RCTs displace other important research (2023, p. 7)

Notably, many political economy factors might help to explain this *methodological fetishism,* a term coined by Poulis and Kastanakis (2020) to describe academic gatekeepers' 'insistence on methodological sophistication and indoctrinated thinking at the expense of the importance, novelty and interestingness of findings' (p. 677). For instance, some authors have argued how, and in contrast with the value and theory-free discourse endorsed by experimentalists, RCTs have clear ideological commitments aligned with dominant (neo)liberal economic reforms (Klees & Edwards, 2014; Klees et al., 2020, Kvangraven, 2020, Steiner-Khamsi, 2012, Tikly, 2015, Wash, 2020). Similarly, comparative education scholars have documented longstanding efforts by international organisations to promote the credibility of their data and methods and, therefore, to bolster the influence of their findings (Edwards Jr et al., 2020, 2023; Heyneman, 2003; Steiner-Khamsi, 2012; Verger et al., 2019).

Abdelghafour's (2017) exploration into the *worm wars* is a case in point to illustrate this phenomenon. The controversy documented by the author started when epidemiologists of the London School of Hygiene and Tropical Medicine failed to replicate, using the original data set, the findings of an influential RCT conducted in Kenya by Miguel and Kremer (2004) on deworming children to improve school attendance. According to Abdelghafour (2017), drawing attention to the failure to replicate results threatened the vested interests of organisations advocating for children's rights who had already built an institutional ecosystem to promote deworming efforts internationally based on the original experimental study. Although the revised estimates, published in the *International Journal of Epidemiology*, followed strict statistical protocols and academic standards, their authors received fierce critique. Some influential voices in the field even condemned the new study for having emerged. It is along those lines that the author of the study points out that:

> . . . [t]he 'worm wars' raise the question of the dynamics of evidence. Initially, the legitimacy of deworming laid exclusively in the scientific credit of [an] RCT. But then, this piece of evidence made its own way (. . .). The construction of large and complex sociotechnical networks transforming evidence into policy eventually makes poverty reduction interventions less sensitive to counter-evidence, and goes against the trial-and-error spirit promoted by RCT advocates (p. 257)

The academic work of one of the authors of this article has also extensively delved into this same issue. An illustrative case in point is the examinations of knowledge production surrounding charter schools in Colombia (Edwards et al., 2020). As elucidated in the analysis in that particular study, to the extent that such policy initiative concurs with the policy framework preferences of organisations such as the World Bank, it becomes evident how

> strategically selective individuals working in organisational contexts produce impact evaluations [like RCTs] that are then disseminated (. . .) in various other publications and fora (. . .) that further enhance the likelihood that the claims being made become part of the discursive context of the global education policy field (p. 144)

Therefore, much like the worm wars example, this case is also salient in highlighting the critical role of expert networks in the diffusion of specific knowledge production and how they might emerge to protect vested interests in global education policy debates (Verger et al., 2019).

Closing remarks: towards critical consumers of education research

Based on the approach of immanent critique, this paper first offered a critique from *within* the RCT narrative. In practice, we argued that one fundamental problem with RCTs lies in the impossibility of building treatment and control groups that remain statistically comparable over time, affecting core assumptions related to the interpretation of impact estimates from experimental evaluations. Moreover, RCTs confront real-life situations (e.g. attrition and the non-passivity of beneficiaries as agents of change in educational interventions) that threaten the ability to safeguard internal and external validity. In principle, experimental evaluators might attempt to address some of the mentioned challenges using statistical techniques to re-balance the mean characteristics of treatment and control groups. However, all these exercises will require subjective judgments regarding which variables or indicators to use in those re-balancing exercises, which contradicts the theory-free and value-free rhetoric that has granted RCTs the status of being the gold standard in evaluation.

In the last section of the article, we shifted to a political economy perspective to reflect on RCTs' privileged status in evidence-based education despite their evident technical and conceptual flaws and limitations. We suggested that the existence of strong (in organisational and financial terms) networks that sponsor and protect knowledge production dynamics helps explain the legitimacy and credibility granted to RCTs, despite their methodological shortcomings. We further argued that the privileged status of RCTs – and their ideological commitment to specific economic ideas – not only discourages the use of other methods but also directs attention away from transforming the political-economic structures that reproduce inequality in low-income contexts, in both the Global North and the Global South.

While other scholars in education have critiqued RCTs, our contribution is the presentation of an immanent critique and its combination with reflections on the vested interests that continue to support knowledge production and mobilisation. The fact that many practitioners, policymakers, and various policy-related organisations continue to advocate for experimental evaluations despite their limitations in practice suggests, ironically, that their usage to inform evidence-based education debates is not at all value-free. In the end, just as RCT proponents make a choice to support these methods, so too do educators, policymakers, fellow researchers, and others have a choice. The choice is to accept RCT results at face value or to critically reflect on their limitations and what (if anything) can be gleaned from their findings. We hope that this paper encourages more of the latter.

Notes

1. The question of how large is large enough is subject to debate in the literature. RCT practitioners usually attempt to solve this question by using *power calculations*, a process

that, provided other statistical assumptions, computes a sample size with a 'specified probability of declaring as significant a particular difference or effect' (Johnson, 1999, p. 767).

2. As Banerjee and Duflo (2009), Nobel laureates in Economic Sciences for their contributions to experimental methods in development economics, put it, '[the] fact that the basic experimental results (. . .) *do not depend on the theory for their identification* means that *a "clean" test of theory* (i.e. a test that does not rely on other theories too) *may be possible'* (p. 172, *emphasis added*).
3. It is worth mentioning that RCT users often resort to supplementary techniques to complement matching exercises to minimise selection bias created by attrition. These include, for example, intention to treat analysis, last observation carried forward analysis, multiple imputations or analyses of the worst-case scenario (Negida, 2017). However, these all rely on assumptions derived from observed variables. In response, Deaton (2010) asserts that '[t]here is nothing wrong with such fixes in principle (. . .) but their application takes us out of the world of ideal RCTs and back into the world of everyday econometrics and statistics' (p. 447).
4. From a statistical perspective, we refer to the similarity in the distributions of residuals. Residuals represent the disparity between anticipated outcomes based on the treatment (e.g. the expected impact of an intervention on school performance) and the actual results observed in the data. In simpler terms, residuals can be considered as the difference between econometric predictions and the actual outcomes, meaning that they contain information from all the non-observed factors (omitted from the statistical model) that might also affect the observed result.
5. We recognize that non-parametric tests, instead of t-tests, might help solve some of the problems described below. The problem with this possible solution to the problem is that using non-parametric tests still entails assuming non-normality in the distribution of errors in econometric models which contradicts the assumption underlying randomisation.
6. Tikly's (2015) reflection on the underlying foundations of the teaching excellence discourse also illustrates this point: 'In empiricist accounts [as represented by RCT studies] are often implicitly informed by a range of normative assumptions despite claims to the contrary. For example, although the evidence relating to the use of incentives and performance-related pay for teachers is mixed, research into the use of [payment-based] incentives to improve teachers' performance is encouraged (. . .) despite a range of evidence that improved teacher motivation is more likely the outcome of multiple causes arising from a number of structural conditions within education systems' (p. 243).

Acknowledgement

We thank Steve Klees, Hikaru Komatsu and anonymous reviewers for their valuable feedback on earlier versions of our manuscript. Any remaining errors are solely ours.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Juan David Parra is Assistant Professor of the Institute of Education Studies from Universidad del Norte (Colombia). He has participated in several evaluation studies, being one of the precursors of realist evaluation (in education) in Latin America. In 2022 he joined an international consortium led by the University of Notre Dame as a technical advisor of USAID's Supporting Holistic and Actionable Research in Education (SHARE) initiative to advance education learning priorities in low-middle income countries.

D. Brent Edwards Jr. is Professor and Graduate Chair in the Department of Educational Foundations at the University of Hawaii. He has published widely on the global governance of education, education policy and political economy, focusing on middle- low-income countries. He is the principal investigator for a three-year $913,000 project funded by the Dubai Cares Foundation entitled, "Crisis Management for Disaster Risk Reduction in Education Systems: Learning from the Elaboration and Integration of Technology-Focused Strategies in El Salvador, Honduras, and Colombia."

ORCID

Juan David Parra http://orcid.org/0000-0003-1902-7660
D. Brent Edwards Jr http://orcid.org/0000-0003-3955-9525

References

Abdelghafour, N. (2017). Randomized controlled experiments to end poverty? *Anthropologie & développement*, *46-47*(46–47), Article 46–47. https://doi.org/10.4000/anthropodev.611

Anderson, K., Zamarro, G., Steele, J., & Miller, T. (2021). Comparing performance of methods to deal with differential attrition in randomized experimental evaluations. *Evaluation Review*, *45* (1–2), 70–104. https://doi.org/10.1177/0193841X211034363

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679. https://doi.org/10.1002/sim.6607

Baker, J. L. (2000). *Evaluating the impact of development projects on poverty: A handbook for practitioners*. World Bank Publications.

Banerjee, A. V., & Duflo, E. (2009). The experimental approach to development economics. *Annual Review of Economics*, *1*(1), 151–178. https://doi.org/10.1146/annurev.economics.050708.143235

Barrenechea, I., Beech, J., & Rivas, A. (2022). How can education systems improve? A systematic literature review. *Journal of Educational Change*, *24*(3), 479–499. https://doi.org/10.1007/s10833-022-09453-7

Barrett, C. B., & Carter, M. R. (2010). The power and pitfalls of experiments in development economics: Some non-random reflections. *Applied Economic Perspectives and Policy*, *32*(4), 515–548. https://doi.org/10.1093/aepp/ppq023

Barrett, C. B., & Carter, M. R. (2020). Finding our balance? Revisiting the randomization revolution in development economics ten years further on. *World Development*, *127*, 104789. https://doi.org/10.1016/j.worlddev.2019.104789

Barry, A. E. (2005). How attrition impacts the internal and external validity of longitudinal research. *Journal of School Health*, *75*(7), 267–270. https://doi.org/10.1111/j.1746-1561.2005.00035.x

Berman, E. P. (2022). *Thinking like an Economist: How efficiency replaced equality in U.S. Public policy*. *En thinking like an Economist*. Princeton University Press.

Bhaskar, R. (2016). *Enlightened common sense: The philosophy of critical realism*. Routledge.

Bhatt, A. (2010). Evolution of clinical research: A history before and beyond James Lind. *Perspectives in Clinical Research*, *1*(1), 6–10. https://doi.org/10.4103/2229-3485.71839

Biesta, G. (2007). Why "what works" won't work: Evidence-based practice and the democratic deficit in educational research. *Educational Theory*, *57*(1), 1–22. https://doi.org/10.1111/j.1741-5446.2006.00241.x

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48* (3), 378–399. https://doi.org/10.17763/haer.48.3.t490261645281841

Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: A systematic review of randomised controlled trials in education research 1980–2016. *Educational Research*, *60*(3), 276–291. https://doi.org/10.1080/00131881.2018.1493353

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, *48*(2), 424–455. https://doi.org/10.1257/jel.48.2.424

Deaton, A. (2020). *Randomization in the tropics revisited: A theme and eleven variations*. National Bureau of Economic Research.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Edwards, D. B., Jr. (2018). *Global education policy, impact evaluations, and alternatives: The political economy of knowledge production*. Palgrave MacMillan.

Edwards, B., Jr., Caravaca, A., Rappeport, A., & Sperduti, V. (2023). The influence of the World Bank on policy formation in education: A systematic review of the literature. *Review of Educational Research*. https://doi.org/10.3102/00346543231194725

Edwards, B., Jr., Morrison, J. & Hall, S. (2020). The suspect statistics of best practices: A triple critique of knowledge production and mobilisation in the global education policy field. *Globalisation, Societies & Education*, *18*(2), 125–148. https://doi.org/10.1080/14767724.2019.1689489

Evans, D. K., & Popova, A. (2016). What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. *The World Bank Research Observer*, *31*(2), 242–270. https://doi.org/10.1093/wbro/lkw004

Fashing, J., & Goertzel, T. (1981). The myth of the normal curve a theoretical critique and examination of its role in teaching and research. *Humanity & Society*, *5*(1), 14–31. https://doi.org/10.1177/016059768100500103

Gale, T. (2017). *What's not to like about RCTs in education? En mobilising Teacher researchers*. Routledge.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. (2016). *Impact evaluation in practice*. World Bank Publications.

Greenberg, D., & Barnow, B. S. (2014). Flaws in evaluations of social programs: Illustrations from randomized controlled trials. *Evaluation Review*, *38*(5), 359–387. https://doi.org/10.1177/0193841X14545782

Hartwig, M. (2008). Introduction. In R. Bhaskar, Eds. *A realist theory of science* (pp. ix–xxv). Routledge.

Henneberger, A. K., Rose, B. A., Feng, Y., Johnson, T., Register, B., Stapleton, L. M., Sweet, T., & Woolley, M. E. (2023). Estimating Student attrition in school-based prevention studies: Guidance from state longitudinal data in Maryland. *Prevention Science*, *24*(5), 1035–1045. https://doi.org/10.1007/s11121-023-01533-1

Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2013). Randomized trials analyzed as observational studies. *Annals of Internal Medicine*, *159*(8), 560–562. https://doi.org/10.7326/0003-4819-159-8-201310150-00709

Heyneman, S. P. (2003). The history and problems in the making of education policy at the World Bank 1960–2000. *International Journal of Educational Development*, *23*(3), 315–337. https://doi.org/10.1016/S0738-0593(02)00053-6

Hubbard, R., Haig, B. D., & Parsa, R. A. (2019). The limited role of formal statistical inference in scientific inference. *The American Statistician*, *73*(sup1), 91–98. https://doi.org/10.1080/00031305.2018.1464947

Isaksen, K. R. (2018). Without foundation or neutral standpoint: Using immanent critique to guide a literature review. *Journal of Critical Realism*, *17*(2), 97–117. https://doi.org/10.1080/14767430.2018.1427180

Johnson, D. H. (1999). The insignificance of statistical significance testing. *The Journal of Wildlife Management*, *63*(3), 763–772. https://doi.org/10.2307/3802789

Joyce, K. E., & Cartwright, N. (2018). Meeting our standards for educational justice: Doing our best with the evidence. *Theory & Research in Education*, *16*(1), 3–22. https://doi.org/10.1177/1477878518756565

Joyce, K. E., & Cartwright, N. (2020). Bridging the gap between research and practice: Predicting what will work locally. *American Educational Research Journal*, *57*(3), 1045–1082. https://doi.org/10.3102/0002831219866687

Kaplan, A., Cromley, J., Perez, T., Dai, T., Mara, K., & Balsai, M. (2020). The role of context in educational RCT findings: A call to redefine "evidence-based practice". *Educational Researcher*, *49*(4), 285–288. https://doi.org/10.3102/0013189X20921862

Khera, R. (2023). Some questions of ethics in randomized controlled trials. *Review of Development Economics*. https://doi.org/10.1111/rode.12996

Kim, T. K., & Park, J. H. (2019). More about the basic assumptions of t-test: Normality and sample size. *Korean Journal of Anesthesiology*, *72*(4), 331–335. https://doi.org/10.4097/kja.d.18.00292

Klees, S., & Edwards, D. B., Jr. (2014). Knowledge production and technologies of governance. In T. Fenwick, E. Mangez, & J. Ozga (Eds.), *World yearbook of education* (pp. 31–43). Routledge.

Klees, S. J., Ginsburg, M., Anwar, H., Robbins, M. B., Bloom, H., Busacca, C., Corwith, A., Decoster, B., Fiore, A., & Gasior, S. (2020). The World Bank's SABER: A critical analysis. *Comparative Education Review*, *64*(1), 46–65. https://doi.org/10.1086/706757

Kumar, S. M. (2016). RCTs for better policy? The case of public systems in developing countries. *Economia Politica*, *33*(1), 83–98. https://doi.org/10.1007/s40888-016-0027-1

Kvangraven, I. H. (2020). Nobel Rebels in disguise—assessing the rise and rule of the randomistas. *Review of Political Economy*, *32*(3), 305–341. https://doi.org/10.1080/09538259.2020.1810886

Leamer, E. E. (2010). Tantalus on the Road to Asymptopia. *Journal of Economic Perspectives*, *24*(2), 31–46. https://doi.org/10.1257/jep.24.2.31

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

Maxwell, J. A. (2020). The value of qualitative inquiry for public policy. *Qualitative Inquiry*, *26*(2), 177–186. https://doi.org/10.1177/1077800419857093

McKnight, L., & Morgan, A. (2019). A broken paradigm? What education needs to learn from evidence-based medicine. *Journal of Education Policy*, *35*(5), 648–664. https://doi.org/10.1080/02680939.2019.1578902

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, *72*(1), 159–217. https://doi.org/10.1111/j.1468-0262.2004.00481.x

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, *40*(2), 1263–1282. https://doi.org/10.1214/12-AOS1008

Muller, S. M. (2020). The implications of a fundamental contradiction in advocating randomized trials for policy. *World Development*, *127*, 104831. https://doi.org/10.1016/j.worlddev.2019.104831

Negida, A. (2017, April). Attrition bias in randomized controlled trials. https://s4be.cochrane.org/blog/2017/02/13/attrition-bias-randomized-controlled-trials/

Odom, S. L. (2021). Education of students with disabilities, science, and randomized controlled trials. *Research and Practice for Persons with Severe Disabilities*, *46*(3), 132–145. https://doi.org/10.1177/15407969211032341

O'Mahoney, J., & Vicent, S. (2014). Critical realism as an empirical project: A beginner's guide. In P. E. En & S. Vicent (Eds.), *Studying organizations using critical realism. A practical Guide* (pp. 1–20). Oxford University Press.

Parra, J. D. (2018). People, personal projects and the challenging of social structures: A contribution to the reflection on the challenges of teaching development studies. *Third World Quarterly*, *39*(11), 2188–2202. https://doi.org/10.1080/01436597.2018.1460594

Pearce, W., & Raman, S. (2014). The new randomised controlled trials (RCT) movement in public policy: Challenges of epistemic governance. *Policy Sciences*, *47*(4), 387–402. https://doi.org/10.1007/s11077-014-9208-3

Porter, S., McConnell, T., & Reid, J. (2017). The possibility of critical realist randomised controlled trials. *Trials*, *18*(1), 133. https://doi.org/10.1186/s13063-017-1855-1

Poulis, K., & Kastanakis, M. (2020). On theorizing and methodological fetishism. *European Management Journal*, *38*(5), 676–683. https://doi.org/10.1016/j.emj.2020.06.006

Sedgwick, P. (2015). Randomised controlled trials: Understanding confounding. *BMJ: British Medical Journal*, *351*, h5119. https://doi.org/10.1136/bmj.h5119

Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised effect sizes. *Journal of Education Policy*, *32*(4), 450–466. https://doi.org/10.1080/02680939.2017.1280183

Simpson, A. (2019). The evidential basis of "evidence-based education": An introduction to the special issue. *Educational Research & Evaluation*, *25*(1–2), 1–6. https://doi.org/10.1080/13803611.2019.1617979

Slavin, R. E. (2002). Evidence-Based Education Policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15–21. https://doi.org/10.3102/0013189X031007015

Stein, H., Cunningham, S., & Carmody, P. (2021). The rise of "behavioral man": Randomized controlled trials and the "new" development agenda. *Human Geography*, *14*(1), 62–75. https://doi.org/10.1177/1942778620987068

Steiner-Khamsi, G. (2012). For all by all?: The World Bank's Global Framework for Education. In S. J. Klees, J. Samoff, & N. P. Stromquist (Eds.), The World Bank and education: Critiques and alternatives (pp. 3–20). Sense Publishers.

Surendran, A., & Kumar, A. (2020). Have RCTs brought back the "empirical" into economics? *World Development*, *127*, 104828. https://doi.org/10.1016/j.worlddev.2019.104828

Thomas, V., & Chindarkar, N. (2019). *Economic evaluation of sustainable development*. Springer Nature.

Tikly, L. (2015). What works, for whom, and in what circumstances? Towards a critical realist understanding of learning in international and comparative education. *International Journal of Educational Development*, *40*, 237–249. https://doi.org/10.1016/j.ijedudev.2014.11.008

Van Belle, S., Wong, G., Westhorp, G., Pearson, M., Emmel, N., Manzano, A., & Marchal, B. (2016). Can "realist" randomised controlled trials be genuinely realist? *Trials*, *17*, 313. https://doi.org/10.1186/s13063-016-1407-0

Verger, A., Fontdevila, C., Rogan, R., & Gurney, T. (2019). Manufacturing an illusory consensus? A bibliometric analysis of the international debate on education privatisation. *International Journal of Educational Development*, *64*, 81–95. https://doi.org/10.1186/s13063-016-1407-0

Vivalt, E. (2020). How much can we generalize from impact evaluations? *Journal of the European Economic Association*, *18*(6), 3045–3089. https://doi.org/10.1093/jeea/jvaa019

Walters, P. B. (2009). The politics of science: Battles for scientific authority in the field of education research. In P. B. Walters, A. Lareau, & S. Ranis (Eds.), *Education research on trial* (pp. 27–60). Routledge.

Wash, I. (2020). Interpreting public policy dilemmas: Discourse analytical insights. *Humanities and Social Sciences Communications*, *7*(1), Article 1. https://doi.org/10.1057/s41599-020-00621-9.

Williams, M. J. (2020). *External validity and policy adaptation: From impact evaluation to policy design* (Vol. 35). Oxford University Press.

Wrigley, T., & McCusker, S. (2019). Evidence-based teaching: A simple view of "science". *Educational Research & Evaluation*, *25*(1–2), 110–126. https://doi.org/10.1080/13803611.2019.1617992

Zhang, X., Wu, Y., Ren, P., Liu, X., & Kang, D. (2015). The relationship between external and internal validity of randomized controlled trials: A sample of hypertension trials from China. *Contemporary Clinical Trials Communications*, *1*, 32–38. https://doi.org/10.1016/j.conctc.2015.10.004