Experiment's persistent failure in education inquiry, and why it keeps failing

Thomas, Gary

DOI:
10.1002/berj.3660

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Thomas, G 2020, 'Experiment's persistent failure in education inquiry, and why it keeps failing', *British Educational Research Journal*. https://doi.org/10.1002/berj.3660

BERJ British Educational Research Journal

BERA BRITISH EDUCATIONAL RESEARCH ASSOCIATION

Check for updates

British Educational Research Journal
Vol. ●●, No. ●●, ●● 2020, pp. ●●–●●

DOI: 10.1002/berj.3660

Experiment's persistent failure in education inquiry, and why it keeps failing

Gary Thomas*
University of Birmingham, Birmingham, UK

Natural scientists are relaxed about the multiple forms experiment takes in their various fields. Yet in education we have for many years constrained our notion of experiment. This methodological circumscription has been self-imposed on the grounds that experiment of a particular, well-defined form offers the clearest evidence of a link between cause and effect in assessing the impact of interventions. I challenge the legitimacy of this assertion and further argue that the model of intervene-and-experiment is ineffective and misleading. There is currently emerging a large body of findings from such experiment, and evaluations of these—like similar evaluations from a wave of experiments in the 1960s and 1970s across education and the applied social sciences—mainly concur on how disappointing the findings from this kind of work are. I argue that interventions are found to have largely nugatory consequences because the influence of independent variables is routinely overwhelmed by powerful contextual influences. I discuss the significance and nature of these contextual influences and question the legitimacy of the idea that one can test interventions with formal experiment in education. I conclude that the assertion that formal experiment *can* be fruitfully employed may drive policy in unhelpful directions, as models which may be successful in some circumstances are rejected on the basis of low effectiveness scores, while others in which potential effectiveness is indicated are unproductively imposed where circumstances are unpropitious. I suggest that a more unrestricted interpretation of 'experiment' needs to return to education discourse.

Keywords: experiment; RCTs; method; inquiry

Introduction

Daddy Pig: *We'll start by doing an experiment.*

Peppa: *What's an experiment?*

Daddy Pig: *It's a way to find out something we don't know—like how many children does it take to lift Mme Gazelle.*

Children [all at once]: *One, a hundred, six . . .*

Daddy Pig: *You're all guessing.*

*University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. Email: g.thomas.3@bham.ac.uk

Danny Dog: *What's the answer?*

Daddy Pig: *I don't know . . . but we can use an experiment to find out. Who wants to try to lift Mme Gazelle?*

Peppa: *Me!* [Tries to lift Mme Gazelle] *I can't lift her.*

Daddy Pig: *Let's try two children*. [Two try but they can't lift Mme Gazelle] *Let's try three children*. [Mme Gazelle rises]

When Peppa asks '*What's an experiment?*', Daddy Pig offers an unusually clear answer. I say 'unusually clear' rather than just 'clear' because the deliberation around methods of experimentation in social science has, I feel, offered anything but clarity. My position is that social science's discourse has obfuscated, not helped, the way to better, more informative inquiry in education research: it has distracted us from our purpose as inquirers.

In this article, I argue that education, and, indeed, the social sciences generally, have constructed an unhelpful notion of experiment. Since the 1920s a particular conception of experiment has crystallised in the social sciences, which is distinct from the more informal, relaxed and fluid idea of experiment embraced in the natural sciences. We might call it the Fisher–Campbell–Stanley notion of experiment (after the major figures in its inception), involving the formal comparison of groups using strictly set methodological ground-rules.

Why did education proceed along this path of supposing that experimentation was obliged to assume a rigidly defined form, and what were the consequences of following this route? I offer a brief history. . .

Experiment in education: the interwar years

Early fondness for what came to be known as 'experimental design' was, in large part, down to the intellectual zeitgeist prevailing when the social sciences were trying to establish their credibility as sciences in the early twentieth century. Out of a desire to achieve respectability in a world where the natural sciences were realising extraordinary feats, and where social sciences, notably psychology, were modestly successful in emulating them (at least in laboratory settings), educators sought what they assumed to be similar methods of verification for their theories. In 1923, McCall published *How to experiment in education*, which drew on, as Campbell and Stanley (1963: 2) put it: '. . . a wave of enthusiasm for experimentation [which] dominated the field of education in the Thorndike era, perhaps reaching its apex in the 1920s'. The contemporaneous work of Fisher (1925) offered the opportunity to enhance the credibility of experimental findings with statistics.

But the enthusiasm for work using experiment proved to be short-lived. As Campbell and Stanley remind us, commentators such as Good and Scates (1954: 716–721) documented a subsequent wave of pessimism that led even staunch advocates of experimentation such as Monroe (1938) to conclude that 'the direct contributions from controlled experimentation have been disappointing'.

Campbell and Stanley attributed the disappointing findings concerning experiment in the interwar years to a number of factors, the principal of which was the lack of sophistication in thinking about experimental design. This lack of sophistication prompted them to write their *Experimental and quasi-experimental designs for research*, in which they taxonomised experimental design, delineating 'pre-experimental designs'—which were taken to be of 'almost no scientific value' (Campbell & Stanley, 1963: 6)—from 'true experimental designs', 'quasi-experimental designs' and others.

Experiment's second coming: the 1960s and beyond

With Campbell and Stanley's opus, the ground-rules had been set for experimentation in education for the second half of the twentieth century. Taxonomy moved to hierarchy, enabled and encouraged by use of qualifiers for experiment such as 'true' and 'quasi', unhelpfully eliding vernacular and technical meanings. 'Experiment' came to mean a formal comparison of groups in the Fisherian tradition, and with the supposed clarification of the form that came with Campbell and Stanley's exegesis, new enthusiasm was found for experiment, which led to a renewed wave of large-scale experimentation in education in the 1960s and 1970s. We might call this experimentation's second coming.

But the new wave quite soon led to a reprise of the pessimism of earlier years, as findings from experiments in natural settings again emerged with disappointing results. These setbacks led two of the best-known researchers in this body of work, Gene Glass and Gregory Camilli, ultimately to make this uncompromising appraisal of experimentation in education during the 1960s and 1970s: '. . . the deficiencies of quantitative, experimental evaluation are thorough and irreparable' (Glass & Camilli, 1981: 23).

Glass and Camilli's comments came after experiment-based evaluations of major intervention programmes such as Head Start (Cicirelli & Associates, 1969), Follow Through (House *et al.*, 1978; Stebbins *et al.*, 1978) and Title 1 (Wargo *et al.*, 1972) had concluded that there were few positive effects from these programmes.

Glass and Camilli were not alone in their assessment of the culpability of experiment in what many took to be inaccurate and misleading appraisals of the value of early years and other compensatory programmes. The failures led distinguished quantitative scientist Lee Cronbach (1975) and others to recommend that evaluators move away from experiment (that is to say, experiment in the Fisher–Campbell–Stanley tradition) and toward ethnographic methods. Rossi (1987) drew similar conclusions about the experimental assessment of social intervention programmes generally during this period—from staff retraining, to prisoner rehabilitation, to those in education.

Experiment's third coming: the discovery of 'evidence'

Despite the clearly articulated warnings from leading experimenters from this second tranche of experimentation in education in the 1960s and 1970s, experiment has had yet another reprise: a third coming. The third coming surfaced out of the call for what came to be known as 'evidence-based practice' (Thomas & Pring, 2004; Parkhurst, 2016) around the turn of the twenty-first century. The new thinking emerged with a

focus on randomisation in experimentation, with the logic for trying yet again after the failures of the interwar tranche and the 1960s/1970s tranche being, as Cook (2001) put it, that 'none of the most heavily criticized studies involved random assignment'. Thus, as time has passed, rather than there being a relaxation of the procedural parameters for the conduct of experiment, they have been tightened with the move to randomised controlled trials (RCTs).

The implication in Cook's comment is that the failure of the post-1960s, second tranche of large-scale experiments could be attributed to lack of randomisation in group assignment. But randomisation adds just one ingredient to the earlier formula —that ingredient being the supposed elimination of allocation bias to groups via randomisation, as if allocation bias had been the wicked problem afflicting the 1960s/1970s tranche of experiments. But allocation bias ought, if it were operating, to have favoured positive findings about intervention. Its elimination, if it were operating, would have led to even more negative findings—a speculation validated by Cheung and Slavin's (2016) finding that effect sizes are significantly higher in non-randomised experiments.

Be that as it may, and notwithstanding major critiques of the logic for randomisation offered by Worrall (2007) and many others, reassurances about the benefits to be garnered from randomisation were trusted by education researchers and policymakers alike. (Critiques of the supposed need for randomisation continue—see Fuller, 2019.) A new fashion for experimentation took hold (Wrigley, 2018: 14 calls it a 'cult'), and with it came new phrases—'evidence-based' and 'What Works' (Thomas, 2021)—and the establishment of national bodies committed to discovering What Works—the What Works Clearinghouse (WWC) in the USA and the Education Endowment Foundation (EEF) in the UK—which have funded hundreds of experiments. This third tranche has now generated enough findings for thoroughgoing evaluations of those findings, and I review the most recent of these evaluations later in this article.

The corollary of the 'evidence-based' and What Works discourse has been a focus on experiment in policy and research funding, accompanied by a renewed taxonomising and ranking of those forms, not just in education but in social policy generally (see Thomas, 2012; Parkhurst, 2016). Indeed, Nutley *et al.* (2013) give 15 examples of hierarchies or ranking systems devised by social scientists for judging the supposed quality of evidence emerging from social research. The taxonomising has had a profound influence on policymakers and practitioners to the extent that some forms of inquiry, which would in many sciences unapologetically be called experiment, are considered in education research to be not even worthy of the name 'research' (see Trybus, 2004 for an example of the influence of this discourse). The hierarchy of experiment forms in education research is revealed in the continued use of terms such as 'true experiment' and 'quasi-experiment', and, with an increasingly doctrinaire tone to the discourse, the anointing of the RCT as the 'gold standard', atop the pyramid (see Hammersley, 2015).

The answer lies in the substrate

Our conception of experiment doesn't have to be like this. 'Experiment' does not have to mean the formal comparison of groups. In natural and applied sciences, in

technology, 'experiment' means to try something out, and from this to infer, and to build the best explanation from one's findings. And in trying something out, one lives in one's environment. One is guided by that environment; one doesn't fight it. 'Right plant, right place!' was the mantra of horticulturalist Beth Chatto, and we'd do well to learn from her advice, for a field of inquiry (or a field of anything, for that matter) must inhabit a substrate that shapes its form.

The nature of the substrate determines the form of analysis. Scientists work with the evidence and the tools at their disposal; they have hunches and theories and they draw inferences about their conjectures on the basis of the evidence they unearth. A palaeoanthropologist, for example, is forced to work with pieces of old bone, DNA fragments, established knowledge from geology, zoology, anatomy, the technology of carbon dating and so on, to build credible accounts of the evolution of *homo sapiens*. Palaeoanthropologists test ideas and build narratives and formulate theory. To do this, they don't need control groups. They work with 'inference to the best explanation'.

My case is that we haven't as a community of inquiry learned the lessons of the past, and the putative improvement and purification of the experiment form is once again leaving us disappointed. Worse, it has complicated and etiolated education research—we always seem to be stretching up for some purer form, to the extent that inquiry is circumscribed, and with that circumscription, distorted and enfeebled. As Daddy Pig recognised, experiment is, in essence, a try-out to discover something we don't know. It involves a hunch, a test and a conclusion in the Popperian tradition of conjecture and refutation (see Popper, 2002). Seen in this way, case study, action research and other forms of inquiry in education should validly claim a right to the descriptor 'experiment'. Indeed, it is to Campbell's credit that while he and Stanley had earlier advised that case studies '... have such a total absence of control as to be of almost no scientific value' (Campbell & Stanley, 1963: 6–7), he later came to the conclusion that case studies could, if conducted with rigour, constitute '... the only route to knowledge' (Campbell, 1988: 377).

As Scriven (in Cook *et al.*, 2010: 109) put it: 'True experiments involve pouring stuff into flasks and finding out whether the result bubbles or turns green ... They have nothing to do with control groups of any kind.' If Scriven's message about methodology is not quite as stark as philosopher of science Paul Feyerabend's '*anything goes*' (Feyerabend, 1993: 4, original emphasis), it is perhaps as simple as 'be eclectic'.

Catachresis and the need for theory, and inference to the best explanation

Near the beginning of the 'evidence-based' movement, Tobin (2005) offered an insight about the intervene-and-experiment model. He suggested that the idea that one could 'scale up' from some experimentally tested prototype should be seen only as a metaphor. One can't 'scale up' in education, as one does in, say, engineering, he argued. He suggested that the mere idea that this is possible is actually less simple metaphor, more catachresis—a profoundly misleading metaphor—conceiving of practitioners as homogenous delivery agents, rather than as co-constructors of change. The conceit that one-off trials can offer What Works prescriptions leads to a

research-informing-practice culture, he suggested, in which schools engage with a rapidly shrinking pool of educational ideas. Or, equally damaging, it may result in good ideas being rejected when they may in fact work very well for certain people in certain circumstances.

This catachresis has its effects at a broad policy level. As Gibbs *et al.* (2011) note, the model of investment in early childhood development has been damagingly rejected by many commentators and policymakers on the back of the supposed 'ineffectiveness' of Head Start and other early years interventions, as 'evidenced' by the intervene-and-experiment trials of the 1960s and 1970s. But the advice to policymakers from this model of evaluation has been shown to be misleading: a major analysis by Hendren and Sprung-Keyser (2019) of compensatory and other spending using new and more sensitive indices to assess the impact of programmes and policies in education and other areas finds that programmes in early childhood education are unequivocally the most effective: 'There is a large "bang for the buck"' (p. 53).

A parallel of this potential to distort influence on policy is seen in research in psychiatry, where, as Bothwell *et al.* (2016) point out, because experiment-based inquiry is more feasible for assessing the therapeutic consequences of psychotropic drugs than it is for assessing the benefits of psychotherapy, the evidence base for drug treatment has become apparently (but only apparently) more robust, actively shaping the nature of the treatment field towards pharmaceutical treatment and away from psychotherapy. Once more, the assessment model distorts the message being received by policymakers.

Elsewhere in social science, in economics, we have witnessed the consequences of pursuing highly complex, quantitative, scientific-looking models which demonstrably (following the 2008 financial crisis) mis-framed and mis-analysed the situation (Lawson, 2009). Often, those models were based on experimental research, and Jackson and Cox (2013) reveal the developing fashion for experimentation in social science in the twenty-first century, showing the quadrupling of articles employing experiment in certain areas over the period since 2000.

Jackson and Cox regret the absence of theoretical grounding on which experiment is based. As they put it: '. . . the result is a lack of coherence, an incomprehensible pattern of small points of light in a dark sky' (Jackson & Cox, 2013: 44). To escape the 'incomprehensible pattern' phenomenon, it is not enough simply to be 'empirical'—to summon up the 'evidence-based' genie (Thomas, 2010). Observations need to be connected in some meaningful theoretical framework, as Pawson and Tilley (1997) and many others have argued. Valid conclusions have to be embedded in a broad range of contextual information that can lead to what is now known by philosophers of science as 'inference to the best explanation' (IBE) (Harman, 1965; Okasha, 2002; Lipton, 2004), to which I shall return in a moment.

Inferential manoeuvres in the dark

Why is there the focus on a particular form of experiment, now manifested in the 'cult' (Wrigley, 2018) for RCTs? This preoccupation with experimental design is replicated nowhere else in the natural or applied sciences. As Haack (2007) put it, science is 'a loose federation of kinds of inquiry' (p. iv), and 'robustness'—a word

often used by claimants of 'evidence-based' methods—is not the preserve of a single form of inquiry. Scientists in varied fields attest to this by using conspicuously diverse methods and techniques to test ideas in their inquiries. But for more than a century, educators and social scientists have remained transfixed with what Parlett and Hamilton (1972) called the 'agricultural-botany paradigm', copying the experiment form of one sliver of scientific enterprise used in plant science, a methodology successfully emulated in pharmacology and medicine.

The logic pursued in those domains has come to be seen as the best and most appropriate one to pursue also in education and some social science, with the assumption of some of the proponents of experiment being that only by looking at comparisons of large numbers of cases in carefully contrived conditions will we be able to attribute cause reliably. But, as Scriven (2008: 22–23) notes, there are many ways, outside the version of the experiment so admired by some social scientists, to go about establishing causation beyond reasonable doubt wherein conclusions are drawn about cause, using straightforward heuristics and reasoning. From the intelligent examination of evidence, theory about cause is built and rejected or refined and ultimately accepted. The process is always about inferring to the best explanation, and it is often the 'science of the singular' (Simons, 1980) that drives forward ideas. I'll return to this in a moment.

What is the USP of experiment?

It is worth looking in a little more detail at what experiments claim to offer us. Here I'll look, as an exemplar, at the proclaimed apogee of the experiment's form, the RCT. Remember that the disappointments of the first two tranches of experimental work in education were put down to, in the first tranche, as Campbell and Stanley (1963) suggested, an inadequate taxonomy of experiment for the guidance of researchers, which meant that researchers were using inappropriate experimental designs, and from the second tranche, to repeat Cook's (2001) comment, 'none of the most heavily criticized studies involved random assignment'. We're now in the third tranche, wherein randomisation is taken to have reduced very substantially the problems of the past.

What, in marketing terms, is the unique selling point (USP) of any kind of experiment, but, for the purpose of argument here, the experiment in its most refined and useful form (in Cook's terms), the RCT? It turns out that the USP is all about being sure about cause and being confident that experiment protocols are adhered to sufficiently strictly to ensure that causal claims are taken seriously. In looking at kinds of research and what they can tell us about causation, philosopher of science Cartwright (2007) argues that the best research methods are those that provide the information you need, reliably, using feasible means and knowing what you know already. She says that experiments such as RCTs may be able to do this in certain circumstances (and I agree that there are limited circumstances in which they may be valuable in education settings—see Thomas, 2016; Morrison, 2020), but may also be very bad at it (see also Phillips, 2019).

Why might they be bad at it? Cartwright runs through the propositional logic of the claims: if the probability of an outcome O (say, an improvement in attainment) is

greater with a putative cause T (perhaps a new teaching method) than without T (no new teaching method) once all 'confounders' (i.e. anything else that could possibly influence a change in attainment) are controlled for, that is sufficient for the claim 'T causes O'. So, paraphrasing her summary of the argument, in a population where 'all other' causes of O are held fixed, any difference in probability of O with T present versus with T absent shows that T (the method) causes O (the improvement in attainment) in that population.

In an experiment, the assumption is that if T causes O in a subpopulation of a given population φ, then T causes O in φ. But of course in a large social science experiment, where impact (of T) is routinely found to be low (see the review of evaluations in the next section), our assumption is that T causes O, to an extent, in at least some members of that population, but—where low impact is found—our conclusion must be that T does not cause O in significant amounts of others, or that T only causes O to a marginal extent overall, or a combination of both. Here is the nub of the issue, summarised nicely by Norman (2003):

> . . . all these [confounding] variables did not just go away at the flip of the allocation coin. They are still there, doing their best to make different people within each group respond differently to the intervention . . . What effects can be identified from such randomised designs are likely to be of such minimal importance as to be of little practical consequence. (p. 582)

Stating that more formally, when we are studying a test population, as Cartwright (2007) puts it:

> To test 'T causes O' in φ via an RCT, we suppose that we study a test population φ all of whose members are governed by the same causal structure, CS, for O and which is described by a probability distribution P. P is defined over the event space {O, T, K_1, K_2, . . ., K_n}, where each K_i is a state description over 'all other' causes of O except T. (p. 12)

Tying this in with Norman's commentary, even in causally homogeneous subpopulations those homogeneous causes are, in a social science frame, likely to be highly influential, attenuating the measured influence of T in causing O. One may be able to say that T causes O, but because in an education landscape K_1, K_2, . . ., K_n (e.g. method, class size, teacher personality and style, catchment, novelty and other factors) exist and are significant in any subpopulation, T will rarely be shown to cause O to any significant extent.

Reviewing experiment-based research

And this is what we find in practice. Not only is the notion of causally homogeneous subpopulations brought into question by conspicuous failures to replicate, as in the notorious failure to replicate the Tennessee STAR impact of class-size findings in California (Bohrnstedt & Stecher, 1999), but the first substantial appraisals of nearly two decades' worth of work from the most recent tranche of What Works experiments are now emerging and showing less-than-encouraging findings. Like the appraisers of the 1960s experiments, today's appraisers find that confounding variables so attenuate treatment effects that those treatment effects prove to be hardly worth finding (see

Viadero, 2009 for an early commentary and Pampaka *et al*., 2016 for a more recent one).

Highly significant are the most recent analyses of experiments funded by major national bodies. Malouf and Taymans (2016), in conducting an analysis of the WWC evidence base on the effectiveness of education interventions, found that most interventions garnered little or no support from technically adequate studies, with intervention effect sizes of minimal magnitude. They say that their findings 'painted a dim picture of the evidence base on education interventions and indicated a need for new approaches, including a re-examination of federal reliance on experimental impact research' (p. 454).

Lortie-Forgues and Inglis (2019) made very similar findings to those of Malouf and Taymans, analysing large-scale experiments commissioned by the UK EEF and the US-based National Center for Educational Evaluation and Regional Assistance. They found the mean effect size of the interventions being evaluated was 0.06 standard deviations, and they concluded that the experimentally evaluated interventions were usually uninformative, with a median Bayes factor of 0.56. Particularly interesting are not only the preponderance of effect sizes around zero, but also the number indicating a negative effect.

There are some positive findings, but Ioannidis (2005), in his classic paper about the use of experimental procedure (principally in medical research), accounts for positive findings as follows: '... if the true effect sizes are very small in a scientific field, this field is likely to be plagued by almost ubiquitous false positive claims' (p. 0698). He proceeds to argue that in any case: '... one would ideally expect all observed effect sizes to vary by chance around the null in the absence of bias' (p. 0700). Working in the field of economics, Young (2018) comes to similar conclusions, suggesting that even those interventions apparently showing an effect may do so merely as statistical and procedural artefacts of experimental procedure. He concludes that: '... many experimental treatments appear to be having no effect on participants' (p. 68).

Commenting on the tranche of experimentation during the 1960s/1970s, Rossi (1987) describes the 'stainless steel law of evaluation', namely that 'the better designed the impact assessment of a social program, the more likely is the resulting estimated net impact to be zero' (p. 4). In particular, the history of special education is replete with examples of experimentally tested special pedagogies, all of which have been characterised by what might be called 'programmatic asymptote and decline' (Thomas, 2009), wherein any initial impact, perhaps attributable to Hawthorne effects, proves to flatten out, ultimately to fall to the *status quo ante* (Thomas & Loxley, 2021).

Lortie-Forgues and Inglis argue that the field needs, as a priority, to understand why experimentally evaluated interventions so often find small and uninformative effects. They offer three possible explanations: (a) because the underlying research on which the intervention is based is unreliable; (b) because the interventions themselves are poorly designed or implemented; or (c) because the interventions in fact are effective but the trials were not designed in such a way as to detect their effects.

Lortie-Forgues and Inglis are astute in their analysis of the nugatory findings here, with point (a) raising issues similar to those Jackson and Cox raised about the need for adequate guiding theory. Point (b) brings to mind Etzioni's (2001) commentary

on the predictive failures of economists, which drew on legendary anthropologist E. E. Evans Pritchard's work on shamanic rainmakers: Evans Pritchard noted that if it failed to rain, the shamans said that either the rain dance was not done right and must be repeated, or that it would indeed rain, but just later than expected. Point (c) gets to the heart of the issue: trials will find it hard to detect effects in the varied environments of education. Where I differ with Lortie-Forgues and Inglis, though, is that they seem to imply that it is the specific form of the trials that is at fault. I suggest that the problem is deeper than this: trials themselves—trials *qua* trials—are not, most of the time, fit for purpose. It is to this issue that I now turn.

Why formal experiments don't work—or, at least, tell us very little—in education

One of the explanations for the conspicuous frailties of traditional quantitative social research rests in the power law principle—more commonly called Pareto's principle, or, in different domains of study, Zipf's law or the 'law of the vital few' (see Newman, 2005). The underlying idea here is that the stability-of-effect assumptions that dominate much social research using experiment are misplaced (Thaler, 1988), and that the influence of particular variables is pervasive, unstable and disproportional, such that they will always overwhelm the influence of others. These influences cannot be dismissed merely as 'noise'. They are a central part of the social landscape and will have their effect not simply by virtue of their value, but via their interaction with other variables, activating or deactivating the potency of other potential determinants of change.

If one accepts the premises of Pareto's principle and its applicability in education, it follows that a few highly significant variables may determine the ultimate effectiveness of most interventions. More than this, the influence of these variables may increase, decay or fluctuate with time, making interaction effects complex and unpredictable. Pareto's principle offers a means of understanding the nugatory and/or short-lived impact of much education innovation, as well as the inadequacies of formal experiment (in the Fisher–Campbell–Stanley tradition) to assess any such impact. If most outcomes are determined by a few factors (such as enthusiasm, sensitivity, engagement, amount of help), then these will always in real-life situations counteract and outweigh the impact of independent variables of interest in experimental research (see Biesta, 2015 for an excellent discussion of quasi-causation). So, any effects of imposing a new teaching method are likely to be occluded by the influence of other variables, such as teacher style, school catchment, parent support, enthusiasm, and so on (the 'K_1, K_2, …, K_n' to which I referred above—the 'causal cakes' of which Cartwright & Hardie, 2012 speak), all working to overwhelm the influence of the variable of interest.

In the worlds of social science, the confounders are bound to win—in Norman's (2003) terms, drowning treatment effects in a sea of unexplained variance. To repeat his warning: the effects that can be identified from experiment 'are likely to be of such minimal importance as to be of little practical consequence'. This is indeed what is found from several decades of experimental research in education ordered or encouraged by governments. As Tom Kane (2016), himself a distinguished contributor of RCT-based research, put it: '… one would have to characterize the past five decades

as a near-complete failure … despite the fact that the National Science Foundation (NSF) and the Institute of Education Sciences (IES) have funded curricula development and efficacy studies for years' (p. 83).

Confounders and how formal experiment deals with them

Confounding dynamics, then, are central to the point I am making about the slightness and/or transitory nature of the consequences emerging from the intervene-and-experiment model in education. Even if one controls for confounders, they are still there—still, in a real-life situation, working to make different people in different places and different circumstances respond differently to an intervention. Equalising the influence of these confounders between or among groups does not mean eliminating their influence. Far from it.

But, highly important though this is, it is separate from the issue of the RCT's claim, as a particular kind of experiment, to *control* for confounders. There is room for confusion here, as the claim to control for confounders may well be read as an elimination of the confounder issue. Clearly, this doesn't follow. But let's put that on a shelf for the moment and look at the claim itself—the claim to control for confounders, since this is at the centre of the experiment's 'offer'.

Discussing ways in which different social inquiry models take account of sources of confounding, Cartwright (2011) says:

> RCTs trust to procedure; other methods import information. Which strategy provides most support for a particular conclusion depends on how confident we can be that the procedures achieve their aim in the case at hand versus the strength of justification for the information imported. (p. 1400)

Experiments claim to control for confounders (this is both their USP and their *raison d'être*), while other methods import information to the analytical arena, interpreting the information from the 'confounders'. Other methods trust the import of information; experiments don't. More than this, experimenters actively *dis*trust it: it is presumed to 'contaminate'. But as Cartwright goes on to point out, for *us* to trust the *experiment*, we must be confident that it is fulfilling its premises (and its promises), its USP, by adequately policing treatment administration via blinding and random assignment, and by using 'techniques—including large sample size—for reliably inferring probabilities from observed frequencies' (p. 1400). Higgins *et al.* (2011) make a summary of these techniques. If it fails to meet these expectations about adequate blinding, sample size calculation and randomisation procedure, we should not take seriously any claim to the robustness of any emerging findings.

Let us look over some of the application problematics which might threaten the fulfilment of those premises when 'true' experiment is used in education:

- Experiments in education are rarely, if ever, appropriately blinded (see Thomas, 2016; Wrigley, 2018). Double blinding is a routine and essential requirement in pharmaceutical trials (Schulz *et al.*, 1995), with a preference for triple blinding. Even simple, single blinding, though, is rarely achievable in education research, with little consideration for the risk of unblinding increasingly considered

significant in medical research (see Bello *et al.*, 2014). In his defence of the methodological benefits of randomisation, Cook (2001) neglects to mention the issues raised by failure to blind, and Connolly *et al.* (2018) do not mention the need for blinding in their review of the place of RCTs in educational research.

- Elimination of allocation bias via randomisation is the *sine qua non* of the 'gold standard' version of the experiment, yet the problems emerging from decay and degradation of randomised samples persist even in medicine, as members of active and control groups behave differently (see Britton *et al.*, 1995; Wood *et al.*, 2004; Vervölgyi, 2011). These problems are inevitably exaggerated in social research, given the agency of the actors. Noncompliance, a problem in pharmaceutical research (Wittes, 2002), is inescapably amplified in social research where active engagement of a mediator such as a teacher, who may or may not 'see sense' in the intervention, is required (Yurkofsky, 2017).
- As Wittes (2002) argues, calculation of sample size in medical trials is complex and depends on event rates of a condition, normal disease progression, expected effectiveness of therapeutic agents and other factors, and will lead to studies involving between several hundred and half a million participants. However, in the trials which are characteristic of much experiment use in education, samples are much smaller. While Connolly *et al.* (2018) suggest that 'Many of these [education trials] have been relatively large-scale', involving over one thousand participants, often individuals are clustered by class or school, effectively making the unit of analysis here the clustered unit, while the primary unit of inference is the individual, raising, as Donner and Klar (2004) explain, issues about the validity of any conclusions drawn. In smaller samples, random allocation does little to attenuate the likelihood of allocation bias, given the increased possibility in such small groups, typically clustered by school or classroom, of random or non-random clustering of favoured (or unfavoured) participants. So the rationale for using randomisation evaporates in these smaller trials, though this provides little discouragement to adherents of the method, who can—but do not necessarily—use complex statistical fixes putatively to correct for the issue (see Campbell *et al.*, 2004).
- Complex experimentation may give the opportunity for the setup (inadvertent or otherwise) of comparisons that favour the 'active' group. The temptation to do this is referred to as 'comparison choice (or control group) bias' (Jadad & Enkin, 2007: 38). Other issues of this kind are discussed by Gueron (2001).
- Subjective assessment of outcome (e.g. whether a child is on- or off-task) is required in experimental appraisals in social research to a far higher extent than in medical or pharmaceutical research. Wood *et al.* (2008) and Oh *et al.* (2019) comment on the distorting consequences of subjective assessment in empirical appraisals, an issue that is amplified if triple blinding is absent.
- There is ample evidence of the equivalent benefits of 'active placebo' effects in social experiments (Moncrieff *et al.*, 1998), bringing to light the likely operation of Hawthorne effects in positive experimental findings, yet active placebos are seldom incorporated into experimental research designs in education. Indeed, the whole notion of control and its intended or unintended placebo effects is a moot one (Howe *et al.*, 2017). As Pawson (2006) puts it, in social research the control 'is not

a piece of apparatus at idle' (p. 51). Karlsson and Bergmark (2015) too have commented on the problematics of 'compared with what?'

- The delivery of education interventions is far more complex and open to variation than the administration of a pill. Koutsouris and Norwich (2018), for example, show how the same 'experimentally tested' intervention was implemented wholly differently in differing local circumstances (see also Goodman *et al.*, 2018). Their findings validate Tobin's (2005) and Biesta's (2015) observations that teachers are co-constructors of change, not identical delivery agents.

Face validity for the experiment, then, comes from the supposed technical robustness of this kind of research, which elevates its status among both lay and expert users of research (Cowen, 2019). It is, however, as vulnerable to distortion as any other kind of research.

Agricultural/botany experiments are not the only fruit

As Cartwright suggests, the agricultural/botany form of the experiment is not the only way of identifying cause. She says that the proclaimed attribute of cause-establishment is sometimes, but only sometimes, a property of the experiment, but this quality is not unique to work using experiment. She gives examples of other forms of cause-establishment, including economists' use of modelling to estimate the degree to which one factor predicts another in a given population ('probabilistic/Granger causality'), where, given the right assumptions, results can legitimately imply causal conclusions.

Scriven (2008: 22–23) concurs. In explaining a process he calls 'general elimination methodology' (GEM), he describes something very similar to IBE. As examples of analysis where GEM is involved, he suggests that the identification of causal route can come from processes as simple as direct critical observation, with direct or simple inductive inference, as in much astronomy, autopsy or engineering breakdown. He also notes the significance of inference based on use of analogy or theory, as in geology. He asserts that inference comes quite validly also from simple direct manipulation, whether it is in the laboratory or the kitchen. Science also makes valid inferences from what he calls 'natural experiments', as in meteorology and epidemiology.

In the simplicity of this process, Scriven coincides with statistician David Freedman (2008), who talks of causal process observations (CPOs) (and in the similarities between CPO, GEM and IBE it seems that the same process has been recognised independently many times over by methodologists, each giving it their own signifier). Freedman stresses the central role of qualitative reasoning and insight in iconic medical breakthroughs.

Okasha (2002) offers two classic case studies of IBE which demonstrate how the form of the inquiry is fashioned to the shape of the inquiry domain. Darwin and Einstein, in their intimations about causation with regard to, respectively, evolution and Brownian motion, each used reasoning about potential explanations, and each offered a more plausible explanation than any which had been offered hitherto. No one said to them, 'That's all very well, but where's your controlled experiment?' Their

explanations explained the world satisfactorily and elegantly, employing evidence and testing ideas in the service of theory.

Even without the control characteristics of agricultural/botany experimentation, it is clear from the examples that Cartwright, Scriven, Freedman and Okasha use that it is possible to draw perfectly valid inferences without the use of controlled experimentation. Much research in the natural sciences is 'singular', using ingenuity, logic and systematic inquiry. But such inquiry is nevertheless, in those sciences, unproblematically seen as 'experiment' (see Thagard, 1998). It involves building a narrative about explanation with perspicacity, intelligence and 'disciplined eclecticism' (Merton, 1976: 169). Its experiment method ('design' is too strong a word) has little in common with the experiment methods used in agriculture and pharmaceutical testing, and embraced formulaically in each of the waves of experimental inquiry in education that I outlined earlier.

To use IBE in any social science, one must make sense of the whole—not by eliminating or controlling for variables of interest, but rather by retaining the fibres that bind an explanatory narrative. Those fibres concern time, place, meaning, intention and much more, all interrelating. We have to use these and combine them with our existing knowledge. IBE is entirely in tune with the realist evaluation of Pawson and Tilley (1997) in asking not 'What works?', but rather 'What works for whom in what circumstances and in what respects, and how?' Whether it's called action research, case study, ethnography or something else, much supposedly non-experiment education inquiry does this kind of thing.

Indeed, much of the most significant and impactful education inquiry of the last 50 years is characterised by the retention of these binding fibres. Look, for example, at Paul Willis's (1993) *Learning to labor*, Harry Wolcott's (1978) *The man in the principal's office* or Stephen Ball's (1981) *Beachside Comprehensive*. In each, the researcher observes the effects of naturally occurring changes, triangulating forms of data collection and forms of analysis to emerge with invaluable insights about the workings of schools. It seems inappropriate to me to exclude these inquiries from the designation 'experiment': in essence, the working methods of these researchers are akin to those of most scientists, whether they be astrophysicists, epidemiologists, meteorologists, palaeoanthropologists or zoologists.

For both the teacher and the natural scientist, using a mix of methods of inquiry—Haack's (2007) 'loose federation of kinds of inquiry' (p. iv)—comes naturally, as long as certain shibboleths of experiment methodology in the Fisher–Campbell–Stanley tradition are relinquished. There is much to be gained from returning to more straightforwardly appropriate ideas of what might constitute an experiment in education inquiry.

Conclusion

I began this article with a quotation from a children's cartoon, offering an ultra-simple commentary on what constitutes an experiment. Ultra-simple, but not simplistic: Daddy Pig's exposition is wholly in tune with contemporary thinking on the nature of scientific inquiry. My point is that there has been an over-development in the sense of what an experiment might be in education. The problem has been that some commentators have succeeded in sequestering understanding of experiment in the

protocols of one narrow branch of experimentation—and by this I don't simply mean RCTs; I mean experimentation understood as the formal comparison of groups (which, of course, includes RCTs). This, I have argued, is illegitimate, asserting validity by affiliating itself with successful inquiry in other, wholly unrelated fields, and at the same time distracting us from what a simple experiment—a systematised conjecture and refutation—can be and what it can do. A preoccupation with what is taken inappropriately to be 'proper' experiment encourages a detour around some of the most potent ingredients of inquiry in education: the reservoirs of knowledge that practitioners have by virtue of their experience.

Scientists are catholic in their attitude to inquiry and experiment—their attitude to inquiry is fluid, flexible, protean. And their work is successful: I gave the example of the methodological eclecticism of palaeoanthropology earlier … and huge strides have been made in palaeoanthropology and in most sciences in recent years. But few would claim similar strides in education science. Indeed, the evaluations reviewed here from the most recent wave of work using experiment to discover What Works mirror those from the 1960s/1970s and the interwar years. They suggest that the contemporary incarnation of this genre of work has delivered very little, with increasing concern on both sides of the Atlantic about the apparent ineffectiveness of experimentally assessed interventions. As economist Young (2018) put it of the disappointing findings of the intervene-and-experiment paradigm in a parallel applied social science field: 'The fact that in so many cases there do not appear to be any (at least) statistically significant effects is, in many respects, much more stimulating than the confirmation of pre-existing beliefs' (p. 68).

Randomisation, the proffered solution to the problems of the 1960s/1970s tranche of experiments, in reality provides no meaningful change. My contention is that what I have called the 'third wave' of *randomised* experiments—just like the second, mainly *un*randomised wave in the 1960s and 1970s—is yielding 'small and uninformative effects' (Lortie-Forgues & Inglis, 2019) because this model of intervene-and-test using the protocols of a particular kind of experiment is, in education, flawed, unable to meet this branch of experimentation's own design expectations and unwilling to take seriously the significance of confounders which vitiate the legitimacy of its findings. This flawed rendition of experiment is giving misleading policy advice as models for change which may be successful in some circumstances are rejected on the basis of low effectiveness scores, while others in which potential effectiveness is indicated are unproductively imposed where circumstances are unpropitious.

I suggest that a more unrestricted interpretation of 'experiment' needs to return to education discourse.

Acknowledgements

I would like to thank Michelle V. Jackson of Stanford University and Tom Perry of the University of Birmingham for their helpful advice on a draft of this article.

Ethical guidelines

This is a theoretical piece in which no subjects and no empirical work with people or animals were involved. Ethics approval was thus not required.

Conflict of interest

There were no conflicts of interest in the production of this article.

Data availability statement

References

Ball, S. (1981) *Beachside Comprehensive: A case-study of secondary schooling* (Cambridge, Cambridge University Press).

Bello, S., Moustgaard, H. & Hróbjartsson, A. (2014) The risk of unblinding was infrequently and incompletely reported in 300 randomized clinical trial publications, *Journal of Clinical Epidemiology*, 67(10), 1059–1069.

Biesta, G. (2015) Improving education through research? From effectiveness, causality and technology to purpose, complexity and culture, *Policy Futures in Education*, 14(2), 194–210.

Bohrnstedt, G. W. & Stecher, B. M. (Eds) (1999) *Class-size reduction in California: Early evaluation findings, 1996–1998 (CSR Research Consortium, Year 1 Evaluation Report)* (Palo Alto, CA, American Institutes for Research).

Bothwell, L. E., Greene, J. A., Podolsky, S. H. & Jones, D. S. (2016) Assessing the gold standard—lessons from the history of RCTs, *New England Journal of Medicine*, 374, 2175–2181.

Britton, A., Murray, D., Bulstrode, C., McPherson, K. & Denham, R. (1995) Loss to follow-up: Does it matter?, *Lancet*, 345(8963), 1511–1512.

Campbell, D. T. (1988) *Methodology and epistemology for social sciences: Selected papers* (Chicago, IL, Chicago University Press).

Campbell, D. T. & Stanley, J. C. (1963) *Experimental and quasi-experimental designs for research* (Boston, MA, Houghton Mifflin Co.).

Campbell, M. K., Elbourne, D. R., Altman, D. G. & the CONSORT group (2004) CONSORT statement: Extension to cluster randomised trials, *British Medical Journal*, 328, 702–708.

Cartwright, N. C. (2007) Are RCTs the gold standard?, *BioSocieties*, 2, 11–20.

Cartwright, N. C. (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).

Cheung, A. & Slavin, R. E. (2016) How methodological features of research studies affect effect sizes, *Educational Researcher*, 45(5), 283–292.

Cicirelli, V. G. & Associates (1969) *The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development (vols 1 and 2). A report to the Office of Economic Opportunity* (Athens, OH, Ohio University and Westinghouse Learning Corporation).

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.

Cook, T. D. (2001) *Reappraising the arguments against randomized experiments in education: An analysis of the culture of evaluation in American schools of education* (Chicago, IL, Northwestern University).

Cook, T. D., Scriven, M., Coryn, C. L. & Evergreen, S. D. (2010) Contemporary thinking about causation in evaluation: A dialogue with Tom Cook and Michael Scriven, *American Journal of Evaluation*, 31(1), 105–117.

Cowen, N. (2019) For whom does 'What Works' work? The political economy of evidence-based education, *Educational Research and Evaluation*, 25(1–2), 81–98.

Cronbach, L. J. (1975) Beyond the two disciplines of scientific psychology, *American Psychologist*, 30(2), 116–127.

Donner, A. & Klar, N. (2004) Pitfalls of and controversies in cluster randomization trials, *American Journal of Public Health*, 94(3), 416–422.

Etzioni, A. (2001, August 13) Economists fail as forecasters, *USA Today*, p. 12A.
Feyerabend, P. (1993) *Against method* (3rd edn) (London, Verso/New Left Books).
Fisher, R. A. (1925) *Statistical methods for research workers* (London, Oliver & Boyd).
Freedman, D. A. (2008) On types of scientific inquiry: The role of qualitative reasoning, in: J. M. Box-Steffensmeier, H. E. Brady & D. Collier (Eds) *The Oxford handbook of political methodology* (Oxford, Oxford University Press), 300–318.
Fuller, J. (2019) The confounding question of confounding causes in randomized trials, *British Journal for the Philosophy of Science*, 70(3), 901–926.
Gibbs, C., Ludwig, J. & Miller, D. L. (2011) *Does Head Start do any lasting good?* Working Paper 17452 (Cambridge, MA, NBER).
Glass, G. V. & Camilli, G. A. (1981) *The future of 'Follow Through'*. Viewpoints 120 (Washington, D.C., National Institute of Education).
Good, C. V. & Scates, D. E. (1954) *Methods of research* (New York, Appleton-Century-Crofts).
Goodman, L. A., Epstein, D. & Sullivan, C. M. (2018) Beyond the RCT: Integrating rigor and relevance to evaluate the outcomes of domestic violence programs, *American Journal of Evaluation*, 39(1), 58–70.
Gueron, J. M. (2001) The politics of random assignment: Implementing studies and affecting policy, in: R. Mosteller (Ed.) *Evidence matters: Randomized trials in education research* (Washington, D.C., Brookings Institution Press), 15–49.
Haack, S. (2007) *Defending science—within reason: Between scientism and cynicism* (New York, Prometheus Books).
Hammersley, M. (2015) Against 'gold standards' in research: On the problem of assessment criteria, paper presented at *Was heißt hier eigentlich 'Evidenz'?*, Frühjahrstagung 2015 des AK Methoden in der Evaluation Gesellschaft für Evaluation (DeGEval), Fakultät für Sozialwissenschaften, Hochschule für Technik und Wirtschaft des Saarlandes, Saarbrücken, Germany. Available online at: www.degeval.de/fileadmin/users/Arbeitskreise/AK_Methoden/Hammersley_Saarbrucken.pdf
Harman, G. (1965) The inference to the best explanation, *Philosophical Review*, 74, 88–95.
Hendren, N. & Sprung-Keyser, B. (2019) *A unified welfare analysis of government policies*. Working Paper 26144 (Cambridge, MA, NBER).
Higgins, J. P., Altman, D. G., Gotzsche, P. C., Juni, P., Moher, D., Oxman, A. D. & Cochrane Bias Methods Group (2011) Cochrane Bias Methods Group (2011) Cochrane Statistical Methods Group: The Cochrane Collaboration's tool for assessing risk of bias in randomised trials, *BMJ*, 343, d5928.
House, E. R., Glass, G. V., McLean, L. & Walker, D. (1978) No simple answer: Critique of the 'Follow Through' evaluation, *Harvard Educational Review*, 48, 128–160.
Howe, L. C., Goyer, J. P. & Crum, A. J. (2017) Harnessing the placebo effect: Exploring the influence of physician characteristics on placebo response, *Health Psychology*, 36(11), 1074–1082.
Ioannidis, J. P. A. (2005) Why most published research findings are false. *PLoS Medicine*, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124
Jackson, M. & Cox, D. R. (2013) The principles of experimental design and their application in sociology, *Annual Review of Sociology*, 39, 27–49.
Jadad, A. R. & Enkin, M. W. (2007) *Randomized controlled trials: Questions, answers and musings* (2nd edn) (Oxford, Wiley-Blackwell).
Kane, T. J. (2016) Connecting to practice: How we can put education research to work, *EducationNext*, 16(2). Available online at: www.educationnext.org/connecting-to-practice-put-education-research-to-work/
Karlsson, P. & Bergmark, A. (2015) Compared with what? An analysis of control-group types in Cochrane and Campbell reviews of psychosocial treatment efficacy with substance use disorders, *Addiction*, 110, 420–428.
Koutsouris, G. & Norwich, B. (2018) What exactly do RCT findings tell us in education research?, *British Educational Research Journal*, 44(6), 939–959.
Lawson, T. (2009) The current economic crisis: Its nature and the course of academic economics, *Cambridge Journal of Economics*, 33, 759–777.

Lipton, P. (2004) *Inference to the best explanation* (2nd edn) (London, Routledge).
Lortie-Forgues, H. & Inglis, M. (2019) Rigorous large-scale educational RCTs are often uninformative: Should we be concerned?, *Educational Researcher*, 48(3), 158–166.
Malouf, D. B. & Taymans, J. M. (2016) Anatomy of an evidence base, *Educational Researcher*, 45 (8), 454–459.
Merton, R. K. (1976) *Sociological ambivalence* (New York, The Free Press).
Moncrieff, J., Wessely, S. & Hardy, R. (1998) Meta-analysis of trials comparing anti-depressants with active placebos, *British Journal of Psychiatry*, 172(3), 227–231.
Monroe, W. S. (1938) General methods: Classroom experimentation, in: G. M. Whipple (Ed.) *Yearbook of the National Society for the Study of Education* (vol. 37, part II) (Bloomington, IL, Public School Publishing Co.), 319–327.
Morrison, K. (2020) *Taming randomised controlled trials in education: A cautionary tale* (London, Routledge).
Newman, M. E. (2005) Power laws, Pareto distributions and Zipf's law, *Contemporary Physics*, 46 (5), 323–351.
Norman, G. (2003) RCT = results confounded and trivial: The perils of grand educational experiments, *Medical Education*, 37(7), 582–584.
Nutley, S. M., Powell, A. E. & Davies, H. T. O. (2013) *What counts as good evidence* (St Andrews, RURU). Available online at: https://research-repository.st-andrews.ac.uk/bitstream/handle/10023/3518/What_Counts_as_Good_Evidence_published_version.pdf?sequence=1
Oh, J. H., Yeatman, S. & Trinitapoli, J. (2019) Data collection as disruption: Insights from a longitudinal study of young adulthood, *American Sociological Review*, 84(4), 634–663.
Okasha, S. (2002) *Philosophy of science* (Oxford, Oxford University Press).
Pampaka, M., Williams, J. & Homer, M. (2016) Is the educational 'What Works' agenda working? Critical methodological developments, *International Journal of Research & Method in Education*, 39(3), 231–236.
Parkhurst, J. (2016) *The politics of evidence: From evidence-based policy to the good governance of evidence* (London, Routledge).
Parlett, M. & Hamilton, D. (1972) *Evaluation as illumination: A new approach to the study of innovatory programs*. Occasional Paper (Edinburgh, Edinburgh University Centre for Research in the Educational Sciences).
Pawson, R. (2006) *Evidence-based policy: A realist perspective* (London, Sage).
Pawson, R. & Tilley, N. (1997) *Realistic evaluation* (London, Sage).
Phillips, D. C. (2019) Evidence of confusion about evidence of causes: Comments on the debate about EBP in education, *Educational Research and Evaluation*, online, 25(1–2), 7–24.
Popper, K. (2002) *Conjectures and refutations: The growth of scientific knowledge* (2nd edn) (London, Routledge).
Rossi, P. (1987) The iron law of evaluation and other metallic roles, in: J. H. Miller & M. Lewis (Eds) *Research in social problems and public policy* (vol. 4) (Greenwich, CT, JAI Press), 3–20.
Schulz, K. F., Chalmers, I., Hayes, R. J. & Altman, D. G. (1995) Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials, *Journal of the American Medical Association*, 273(5), 408–412.
Scriven, M. (2008) A summative evaluation of RCT methodology: An alternative approach to causal research, *Journal of Multidisciplinary Evaluation*, 5, 11–24.
Simons, H. (1980) *Towards a science of the singular* (Norwich, CARE, UEA).
Stebbins, L. B., St. Pierre, R. G., Proper, E. C., Anderson, R. B. & Cerba, T. R. (1978) An evaluation of Follow Through, in: T. D. Cook (Ed.) *Evaluation Studies Review Annual* (vol. 3) (Beverly Hills, CA, Sage), 571–610.
Thagard, P. (1998) Ulcers and bacteria I: Discovery and acceptance, *Studies in History and Philosophy of Science, Part C: Studies in History and Philosophy of Biology and Biomedical Sciences*, 29(1), 107–136.
Thaler, R. H. (1988) Anomalies – the ultimatum game, *Journal of Economic Perspectives*, 2(4), 195–206.
Thomas, G. (2009) 'What Works' as a sublinguistic grunt, with lessons from catachresis, asymptote, football and pharma, *Research Intelligence*, 106, 20–22.

Thomas, G. (2010) Evidence began in 1998, *Research Intelligence*, 109, 14–15.
Thomas, G. (2012) Changing our landscape of inquiry for a new science of education, *Harvard Educational Review*, 82, 26–51.
Thomas, G. (2016) After the gold rush: Questioning the "gold standard" and reappraising the status of experiment and randomized controlled trials in education, *Harvard Educational Review*, 86(3), 390–411.
Thomas, G. (2021) 'Evidence-based' as a zombie idea: 'Evidence' and symbolic power in education discourse, *Oxford Review of Education* (under review).
Thomas, G. & Loxley, A. (2021) *Deconstructing special education and constructing inclusion* (3rd edn) (London, Open University Press).
Thomas, G. & Pring, R. (Eds) (2004) *Evidence-based practice in education* (Maidenhead, Open University Press).
Tobin, J. (2005) Scaling up as catachresis, *International Journal of Research & Method in Education*, 28(1), 23–32.
Trybus, M. (2004) The challenge and hope of scientifically based research, *Viewpoints*, 11, 3–4. Available online at: www.ncrel.org/policy/pubs/html/vp11/essay.htm
Vervölgyi, E., Kromp, M., Skipka, G., Bender, R. & Kaiser, T. (2011) Reporting of loss to follow-up information in randomised controlled trials with time-to-event outcomes: A literature survey, *BMC Medical Research Methodology*, 11(1), 130.
Viadero, D. (2009) "No effects" studies raising eyebrows. *Education Week*, 28(27), 14–15.
Wargo, M. J., Tallmadge, G. K., Michaels, D. D., Lipe, D. & Morris, S. J. (1972) *ESEA Title I: A reanalysis and synthesis of evaluation data from fiscal year 1965 through 1970*. Final Report, Contract No. OEC-0-71-4766 (Palo Alto, CA, American Institutes for Research).
Willis, P. (1993) *Learning to labour* (Aldershot, Ashgate) [first published by Saxon House, 1978].
Wittes, J. (2002) Sample size calculations for randomized controlled trials, *Epidemiologic Reviews*, 24(1), 39–53.
Wolcott, H. (1978) *The man in the principal's office* (Prospect Heights, IL, Waveland Press Inc.).
Wood, A. M., White, I. R. & Thompson, S. G. (2004) Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals, *Clinical Trials*, 1(4), 368–376.
Wood, L., Egger, M., Gluud, L. L., Schulz, K. F., Jüni, P., Altman, D. G. *et al.* (2008) Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study, *British Medical Journal*, 336(7644), 601–605.
Worrall, J. (2007) Why there's no cause to randomize, *The British Journal for the Philosophy of Science*, 58(3), 451–488.
Wrigley, T. (2018) The power of 'evidence': Reliable science or a set of blunt tools?, *British Educational Research Journal*, 44(3), 359–376.
Young, A. (2018) Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results, *The Quarterly Journal of Economics*, 134(2), 557–598.
Yurkofsky, M. M. (2017) *The restructuring of educational organizations: From ceremonial rules to technical ceremonies*. Ph.D. thesis, Harvard Graduate School of Education. Available online at: https://dash.harvard.edu/bitstream/handle/1/33797215/YURKOFSKY-QUALIFYINGPAPER-2017.pdf?sequence=1