Ch. 6 The problem with randomised controlled trials for Education

Terry Wrigley

Abstract

Terry Wrigley engages with Philpott & Poultney's (2018) chapter on randomized control trials (RCTs), particularly the debates around whether they are a useful approach for informing teaching and school development. He critiques as reductionist the claim that RCTs provide a 'scientific' form of research in education. Further, he reflects on the adoption of meta-analyses as a 'gold standard' by various national agencies, and the circumstances which have made Hattie's (2008) Visible Learning publications popular. The chapter looks towards viable alternatives in engaging teachers with research in order to develop the argument for research-informed teacher learning.

Introduction

As a guest at Carey Philpott's inaugural lecture as Professor of Teacher Education at Leeds Beckett University, I soon began to appreciate his powerful combination of an open-minded search for ways of improving educational practice with a critique of superficial thinking and false trails. At the time of his unexpected and untimely death, he was becoming a major figure in the field.

Philpott's guide to *Evidence-Based Teaching* (2018), in collaboration with Val Poultney who completed the book, brought together the increasingly dominant methods of randomised controlled trials and systematic reviews with more collegial professional learning such as learning rounds. It explored potential gains that teachers could make in learning from medical practitioners whilst clearly exposing the differences between the two fields of practice. The book follows the theoretical lead of Hammersley (2015) and Pring and Thomas (2004), among others, but is positioned closer to practice, and was written at a time when RCTs are increasingly being portrayed as the only valid research method.

Philpott & Poultney's (2018) Chapter 3 *Randomised Controlled Trials* explains a number of practical problems including the difficulties of randomisation, inconsistency of implementation and the impossibility of 'double blinding'. I will attempt here to build on some of the most crucial issues which are, in my view, fundamental to understanding why research for teachers' learning cannot be built on such a foundation. Examples will be taken from major projects receiving large amounts of government funding.

RCTs and human agency

Philpott and Poultney (2018) make a sober and knowledgeable assessment of the difficulties of teachers seeking to emulate the role of medical practitioners in drugs trials. Unlike the administration of pills and placebos in a 'double blinded' drugs trials, teachers cannot ask more open or higher-level questions without knowing what they are doing, nor indeed can students remain unaware that something different is going on.

Basically, the 'treatment' in education cannot be carried out effectively without a positive commitment from the teachers, but that immediately introduces a distortion in the trial: the students are responding not only to a particular procedure but to the teacher's enthusiasm, thus introducing a *Hawthorne effect* (Philpott and Poultney 2018: 43). Conversely, teachers who are assigned control groups can easily convey a lack of interest if they feel they are positioned as devaluing certain practices.

In a major randomised control trial in England of Fresh Start, a resource and method to improve the reading of struggling 11 year old readers (Education Endowment Fund, 2015), the treatment group headteachers had chosen to join the programme because they are enthusiastic, and the teachers are explicitly told that passion is a requirement:

The FS-specific teaching style is a core element of this intervention which encompasses teacher's passion, praise for students and a dynamic pace for the lessons. (p5)

This is not just a practical difficulty in one instance, but is central to the use of RCTs in education. As Pawson (2006:27) points out, whereas drugs trials try to eliminate the human factor because 'human volition is seen as a contaminator', social change is brought about precisely *through* the human agent. In other words, RCTs cannot by their very nature accommodate teachers' enthusiasm and passion, which raises questions about the evidence being produced.

This problem of agency is well illustrated by a recent attempt to evaluate project-based learning (Education Endowment Fund, 2016): nearly half the intervention schools dropped out, suggesting a problem in convincing the teachers and also perhaps students that it is worth the effort. It should come as no surprise to those working in the classroom that implementing such pedagogies requires an energetic professional commitment, especially when they break from the tightly controlled pedagogies which have become the norm in a high-stakes accountability regime.

What do control groups actually do?

Philpott and Poultney point to one important difficulty here. They emphasise that the practices being trialled can 'leak' into the control group.

It can occur with students who share experiences with one another. It can occur when a teacher teaches classes (clusters) on both sides of the RCT. It can occur when teachers talk to one another about what they are doing. (p42)

To return to the hypothetical example of asking more open and challenging questions, are the control group teachers going to be asked:

- a) to do what they would normally do? or
- b) to stick rigidly to closed questions without cognitive challenge?

These two options will produce very different effect sizes, according to the distance between the experiences of the treatment and control groups.

Pawson is again worth quoting, for the precision with which he punctures the myth of 'control':

And what of the control? This is not a piece of apparatus at idle. This is not the world in repose. This is no vacuum, because there is no such thing as a policy vacuum. Control groups or control areas are in fact kept very busy. (Pawson 2006:51)

The problem can be quickly illustrated by another large randomised control trial also funded by the Education Endowment Fund in England, the key channel for government-funded research. The trial concerned 'teacher observation', though in fact it was limited to one specific approach which in many ways runs counter to the intended purpose of collegial peer observations to

improve pedagogy. Not only did treatment group teachers fail to carry out as many observations as required, the control group were already committed to more fruitful modes of observation.

Almost three-quarters of the control group schools were already doing some peer observation prior to the intervention. The lack of impact seen in this study may be because the structured Teacher Observation intervention was no more effective than existing practice rather than because general peer observation has no impact. (Worth et al 2017 p3)

In the earlier example of a synthetic-phonics based approach to remedial teaching of reading at the start of secondary school (Education Endowment Fund, 2015), the report gives no explanation of what control groups were doing; nor is there information on whether control group teachers were offered parallel training or additional staffing to reduce group size and put them on a par with the treatment group.

Proofs of causality

Among the claims made by proponents of RCTs in education, Philpott and Poultney (p41) highlight the assertion

that they are the only form of educational research that can identify causality in educational interventions (Torgerson, 2001).

Philpott and Poultney's critique is very firmly that RCTs

cannot, by themselves, tell us the mechanism through which some thing works. So the results of an RCT might show that a particular classroom strategy in mathematics produces improved mathematics scores. However it cannot tell us what the 'active ingredient' is. Which part of the strategy is having the effect and/or how? (Philpott and Poultney 2018: 44)

They add that there are also contextual and psychological variables to consider 'such as gender, social class, ethnicity, identity and personal aspirations' (p.46).

The problem goes deeper, however, and at a philosophical level requires a criticism of *empiricism*. The word 'empiricism' does not mean that observations of reality are unimportant, but rather that we need to dig deeper than surface impressions of phenomena in order to

identify the forces and structures which affect a situation. Bhaskar (1978) makes an important conceptual distinction between the *real* (the underlying forces), the *actual* (what happens), and the *phenomenal* (what we observe). Most natural and social situations are open systems in which different causal forces interact, reinforcing or blocking each other, and also encounter contextual factors in order to *actualise*. For this reason, a force (eg law of nature, social structure) can exist even when it is not effective or observed; conversely, observed regularities are insufficient to prove causality.

Empiricists, in the tradition of David Hume (1711-1776) known for his philosophical empiricism and scepticism, act as if *regularities are all*. Indeed Hume held that it is impossible to move beyond the regularities to get down to implied causes. The procedures used in RCTs deal with regularities but not with causes. Indeed, many designers of RCTs make no attempt to identify deep causes.

There is considerable irony here. Gorard and See (2013:4) argue that a causal model would require not only repeated association, but also a meaningful sequence (X must always precede Y), a measurable relationship (variation in the strength of X must link to variation in Y), and the ability to explain a coherent mechanism. However, in the practical conduct of a major RCT, the same researchers failed even to look for the causal explanations. Thus in their study of a remedial reading programme (Education Endowment Fund, 2015), there is no attempt to discover the reason why the target students are still finding reading difficult, why the program works - if it does - or why it appears to work well with some students but not others.

As Philpott and Poultney rightly say, RCTs generally need to be complemented by various qualitative methods to identify causality, for example close observation accompanied by interviews. Other researchers point to the importance of theory, which includes forming alternative hypotheses and building on accumulated theoretical knowledge. Pawson (2006:47) argues that 'medical treatments... are the embodiment of years of theory-testing... Whole episodes of pure science are played out, and their lessons digested, before the applied science kicks in.' Hammersley (2015:4) points out that 'in the drug field, RCTs are used as a complement to laboratory work, which will have produced a considerable body of knowledge about the drug' whereas 'in social fields RCTs are usually expected to provide the whole scientific knowledge base for the "treatment"'.

We may contrast this with the argument presented by Tom Bennett, popular (and populist) founder of ResearchED, who works entirely from regularities but does not appear to recognise his own trivialisation of theory:

If I apply a Bunsen flame to water, I may be surprised (because I am an idiot) to see it bubble and vanish (let's call it 'boil') when it gets to 100 degrees Celsius. If I propose that this is a routine event, and every time I do the same I obtain the same result, then I can reasonably be said to have a good piece of scientific explanation.

Science normally proceeds on this formula:

- *Form a question: does sound travel faster in water than in air?*
- *Make a hypothesis: yes it does. 'Sound travels faster in water than air.'*
- *Make a prediction: what would I observe if my hypothesis were true? Well, for a start, perhaps I would hear a noise more quickly underwater than I would on land.*
- *Test the prediction: gather evidence to see if the real world behaves the same way as your prediction. Get your flippers on.*
- *Analysis: what does the evidence show? What do we need to do next? and if the evidence proves the hypothesis to be false, what new hypothesis can we suggest?*

(Bennett 2013:21)

Bennett, who insists that RCTs and their synthesis are the only valid 'scientific' form of educational research, appears to regard experiments as a faithful reflection or re-enactment of reality, and portrays the process of designing trials as moving along smoothly from data or casual observations without theory, or indeed anything more than a superficial sequential hypothesis.

This is a problem which permeates much of the work of the Education Endowment Foundation in England, including its 'Teaching and Learning Toolkit' (more later), a reflection of the hegemonisation of the RCT methodology. It also runs through John Hattie's (2008) world-famous *Visible Learning* project. As Danish professor Steen Nepper Larsen, of Aarhus University, expresses it:

John Hattie never explains what the substance of an effect is. What is an effect's ontology, its way of being in the world? This question cannot be answered simply by

providing a number, an 'effect size', as a supposed quantification of the impact of a pedagogic practice on student learning. (Larsen 2013)

Meta-analysis and beyond

Carey Philpott, with Val Poultney (Philpott and Poultney, 2018), clearly recognise the need to make research findings more accessible to teachers, but they make a plea for a more open and intelligent process of reviewing research where decisions do not just depend on technical criteria. Since the technical criteria usually exclude important qualitative approaches, this is extremely welcome. For example, they support Hammersley's argument that 'positivistic or scientific research should not be privileged over the academic-narrative approach to systemic view' (Philpott and Poultney 2018: 29). They reiterate Maclure's (2005:398) argument that systematic review can 'set limits on the ways that the world can be viewed and construed, and establishes what will count as truth'.

Reviewing research must be intelligent and conducted with a deep understanding of the field, particularly when it comes to research-informed teacher learning. It involves decisions as to 'what quality is and what it is not' as well as an understanding of practitioners' perspectives (Philpott and Poultney 2018: 29-30). In Hammersley's words (2001:549, cited Philpott and Poultney 2018: 29), a review of the literature:

can involve judging the validity of the findings and conclusions of particular studies and thinking about how these relate to one another, and how their interrelations can be used to illuminate the field under investigation. This will require the reviewer to draw on his or her tacit knowledge, derived from experience, and to think about the substantive and methodological issues, not just to apply replicable procedures.

Unfortunately, in England and elsewhere, strong governmental pressure (including centralised control of research budgets) is being exerted in favour of the 'meta-analysis' approach which, after excluding many important studies on technical grounds, simply aggregates the 'effect sizes' from multiple RCTs.

Firstly, this represents a misunderstanding of the term 'effect size'. This expression does *not* mean *how effective is this practice* but rather *how well has this RCT, as an experiment, managed to show that this practice has some effect*. In other words, it is an evaluation of research design,

not of teaching methods (see Simpson 2017; 2018). As he demonstrates, effect size can be increased by how an RCT is set up, including:

- ensuring that the control group experience very limited teaching
- researching a more limited population (eg 11-year-old boys with reading difficulties)
- selecting outcome measures which are closely related to the nature of the intervention, rather than more general assessment tools.

Averaging the effect sizes of multiple RCTs simply compounds the problem. Although it can sometimes be appropriate to bring together RCTs which are fairly similar but too small to reach firm conclusions, the mean effect size is dangerously misleading when RCTs are thrown together which are very different - an error commonly known as 'apples and oranges'.

Thus, for example, the meta-analysis on *Feedback* included in the government-funded Education Endowment Fund, Teaching and Learning Toolkit (subsequently abbreviated to *Toolkit*) throws together many disparate studies under the heading 'Feedback', some of which are simply a repetition or clarification of instructions for a classroom task (Wrigley 2018). Confusion is sown by averaging experiments involving detailed marking of completed written work, informal spoken advice directed to individuals while they pursue a practical activity, handing out test scores in the hope of stimulating students to try harder next time, advising students as they design their enquiry, and so on. It is not necessarily the case that 6 year olds will respond to feedback in similar ways to 16 year olds, nor that mathematics benefits from similar forms of feedback to drama. (See also Lilley 2018 for a critique of Hattie's treatment of this same topic.)

In evaluating the impact of sports participation on academic attainment, the Toolkit draws on systematic reviews of research which throw together such diverse activities as cheerleading by girls while the boys actually play the game, and the therapeutic use of yoga and massage to improve concentration. It is not surprising that the results are inconclusive. Averaging effect sizes is not only flawed for the reasons provided by Simpson (see above) but serves as a diversion from examining *why* a particular pedagogical method might impact on learning. The search for 'what works' tells us nothing about *how* it works. This can only be glimpsed by scrutinising the original studies which are hidden behind an average.

In the case of sports participation, one unusually effective intervention involved premier league soccer clubs setting up study centres for underachieving students from local schools.

Participation was only for 20 hours spread across six months, but these young people enjoyed a boost to status and self-esteem through meeting star players and visiting the club's boardroom and museum. The intervention was well resourced, including one-to-one mentoring and dedicated ICT suites. They enjoyed a personalised curriculum geared to their specific needs in literacy, numeracy and ICT skills and designed round practical and situated activities which were meaningful to the students: mathematics trails; counting the seats and measuring the pitch; using gate receipts and sales in the shop, restaurant and kiosks for work on numeracy and data handling; writing match reports; researching and writing player biographies; compiling a sports magazine or match programme; using sports-themed tasks to learn how to search the internet.

An opposite error can be found in Hattie's claim that 'direct instruction' is a superior mode of teaching. In fact the research he is amalgamating is based on one particular 'instructional design' (to use the American terminology) based not on teaching 'from the front' alone but a carefully planned sequence involving learner engagement, modelling, guided practice, monitoring and independent practice / transfer (Hattie 2009: 205-6). Although Hattie insists (pp.208-12) that inquiry methods and problem-based learning are less efficient for learning facts and concepts, he concedes that they are better for engaging students, understanding the principles that link concepts together, longer-term recall, applying knowledge, solving problems, critical thinking and scientific process! Unfortunately the dials which decorate these pages are extremely misleading, since they do not reflect such differences of purpose; the simple meta-analytic averaging of mean effect sizes could easily seduce teachers into discarding inquiry methods.

It is worth noting that Gene Glass, who originated the idea of meta-analysis, has issued a sharp warning about heterogeneity ('apples and oranges'):

Our biggest challenge is to tame the wild variation in our findings not by decreeing this or that set of standard protocols but by describing and accounting for the variability in our findings. The result of a meta-analysis should never be an average; it should be a graph (Robinson 2004:29, my italics).

Meta-meta-analysis

It should be obvious from the discussion so far that large-scale structures built on these methods will have shaky foundations. For convenience, we could call this 'meta-meta-analysis' or perhaps 'mega-synthesis'. The Toolkit (Higgins et al 2014) conducts and draws together meta-analyses under 35 different headings to pronounce on their relative effectiveness in assisting

students disadvantaged by poverty. The mean effect sizes are translated into 'months of additional progress', ranging from +8 to -4 months. Such an approach would appear laughable to medical practitioners: imagine declaring that surgery is more or less effective than painkillers in prolonging life, regardless of whether the problem is a hernia or a headache.

Hattie (2009) claims that his *Visible Learning* project synthesises over 800 meta-analyses involving 50,000 research studies. It is not surprising that this has global appeal to busy teachers, but it is seriously misleading. (See visablelearning.blogspot.com for an ongoing investigation of Hattie's methods and conclusions.)

All of this claims to be a rigorous and 'scientific' approach to educational research, but has little resemblance to the methods of natural science. As argued earlier, it neglects theory and fails to investigate causality. Furthermore, it misapplies techniques from one field of enquiry to another. It thus exemplifies the error known as reductionism, as when biologists seek to explain animal behaviour solely in terms of biochemical change, or psychologists explain human behaviour solely in terms of neurological mapping of the brain. Educators could learn a lot from Hilary and Stephen Rose's remarks about 'biologism', i.e. the misuse of biological explanations to account for psychological and social phenomena, for example when war is explained as a form of animal aggression or human thought by analogy with computer technology. They discuss the real-life consequences of these forms of scientific reductionism, including the use of ritalin, behaviourist punishment regimes, and beliefs in fixed genetically-determined intelligence, all of which blinker practitioners to the complexity of social context and experience.

Research is not just experiments

Scientific experiments, as Critical Realists such as Bhaskar (1978; 1979) explain, are not a simple reflection of reality but a deliberate move to create a closed situation by excluding or stabilising most of the factors operating in naturally open systems. Stephen Rose, as a biologist, does not propose abandoning experiments but places serious warnings of the dangers:

Effective experiments demand the artificial controls imposed by the reductive methodology of the experimenter, but we must never forget that as a consequence they provide at best only a very simplified model, perhaps even a false one, of what happens

in the blooming, buzzing, interactive confusion of life at large, where things rather rarely happen one at a time. (Rose 2005:28)

Further:

What happens in the test-tube may be the same, the opposite of, or bear no relationship at all to what happens in the living cell, still less the living organism in its environment. (p79)

Where they are used, the researcher must make an effort to rebuild the complexity of nature's open systems - and almost every system that occurs in the real world is open. A parallel applies to RCTs as the supposed equivalent of experiments in education, which comes back to questions about the evidence being produced.

They are marred by a failure to recognise key human characteristics such as agency, volition, intentionality, reflection, curiosity, confusion and understanding. They amount to pseudo-science when they fail to ask about purpose - what exactly are our educational aims? what is a particular practice effective *for*? Education is marked by a multiplicity of aims - acquisition of factual knowledge, problem solving, longer-term cognitive development, skilled performance, aesthetic or ethical qualities, socialisation. Tight notions of 'evidence' in the sense of numerical data (effect sizes) risk abandoning such values and aims as world citizenship, multiculturalism, enlightenment, democracy, solidarity, character, virtue, knowledge and Bildung (see Rømer 2014: 115). This is an even greater risk when government insists on 'evidence-based teaching' in a context of high-stakes accountability: the two reinforce each other, and even if the former produces higher scores in the latter, education is diminished.

Education is unavoidably about openness, emergence and consciousness. As Biesta powerfully explains, pedagogical activity involves 'open, recursive, semiotic' systems which linear mechanistic models cannot reflect.

Such conditions can be described as those of closed systems: systems that are in a state of being isolated from their environment. Open systems, on the other hand, are systems that are characterised by a degree of interaction with their environment. Whereas closed systems operate deterministically, open systems operate at most probabilistically. Recursive systems are systems that in some way feed back into themselves, so that the behaviour of the system is the result of a combination of external factors and internal

dynamics. Semiotic systems are systems that do not operate through physical force but through the exchange of meaning. (Biesta 2010:496)

It is not scientific to treat open systems as if they were closed ones, or social situations as if they were biological or physical phenomena.

Concluding remarks

Experiments are not the only form of scientific investigation. Indeed many scientific fields use few of them: think of astronomy, meteorology, evolution - perhaps biology overall. As Gary Thomas pointed out (2004:1-6), experiments are generally used to verify rather than advance knowledge, and many discoveries and inventions (eg penicillin, superconductivity, even aeroplanes) have not arisen from systematic procedures. Scientific method depends heavily on reflective observation, intelligent noticing, trial and error, and even intuition. We should not neglect the different stages involved in scientific enquiry:

If various stages in the employment of evidence are traversed in moving toward knowledge - a bricolage / hunch stage, an inspirational stage, a discovery stage and a corroborative / confirmatory stage - the notion of evidence-based practice focuses on evidence at the confirmatory stage, on the systematic collation of research studies for use by practitioners and policy-makers. (Thomas 2004:10).

In the psychology of human learning an insistence on experimental procedures can lead to disastrous misconceptions, simplification and narrowing (see Vygotsky's 1925 critique of Pavlovian behaviourism). Despite the importance of RCTs in drugs trials and many other forms of discrete intervention in medicine, evidence-based medicine relies on a wider concept of evidence, including the practitioners' experience and empathetic engagement with patients (Greenhalgh et al 2014).

Carey Philpott was centrally concerned with research-informed teacher learning. His many research articles, and the book completed after his sudden death by Valerie Poultney (Philpott and Poultney, 2018), explore a range of forms which this might take. His reaching out and willingness to learn from other fields such as medicine were never uncritical (see for example his discussion of 'rounds' as a form of medical education, Philpott 2017). In his companion piece written with Poultney, he examines the adoption of 'learning rounds' in Scotland, pointing out

the importance of professional agency and collaboration. It is crucial to avoid a mere 'audit', as when teachers are concerned with box-ticking features of teaching rather than a deeper causal enquiry leading to a 'theory of action'.

There are countless examples in the history of education of research which has been practically influential without anything resembling an RCT: think for example of studies by Douglas Barnes and others on classroom language interactions. There is a long international tradition of practitioner research, exemplified by Elliott, Kemmis, Groundwater-Smith and Mockler among others. Carey Philpott's truncated research career belongs to this, because it combines critique, human values, educational aims, and a persistent desire to make young people's learning more successful.

References

- Bennett, T (2013) *Teacher proof: why research in education doesn't always mean what it claims, and what you can do about it*. London: Routledge
- Bhaskar, R (1978) *A realist theory of science*. Hassocks: Harvester Press
- Bhaskar, R (1998) *The possibility of naturalism: A philosophical critique of the contemporary human sciences*. London: Routledge
- Biesta, G (2010) Why 'what works' still won't work: From evidence-based education to value-based education. *Studies in the Philosophy of Education* 29,, pp491-503
- Education Endowment Fund (2015) *Fresh Start: Evaluation report and executive summary* (Gorard, S, Siddiqui, N and See, B H).
[https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Fresh_Start_\(Final\).pdf](https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Fresh_Start_(Final).pdf)
- Education Endowment Fund (2016) *Project-based learning* (The Innovation Unit).
<https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/project-based-learning/>
- Gorard, S and See, B H (2013) *Overcoming disadvantage in education*. London: Routledge
- Greenhalgh, T, Howick, J and Maskrey, N (2014) Evidence based medicine: a movement in crisis? *BMJ* 2014: 348:g3725 (13 June)
- Hammersley, M (2001) On 'systematic' reviews of research literatures: A 'narrative' response to Evans and Benefield. *British Educational Research Journal* 27(5): 544-54
- Hammersley, M (2015) Against 'gold standards' in research: On the problem of assessment criteria. DeGEval conference, Saarbrücken
- Hattie, J (2009) *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge
- Larsen, S (2013) Know thy impact: Blind spots in John Hattie's evidence credo, translated from *Dansk Paedagogisk Tidsskrift* 3, pp69-72

- Lilley, G (2018) Feedback. visiblelearning.blogspot.com/p/feedback.html (accessed 22 Dec 2018)
- Maclure, M (2005) Blarney bordering on stupidity: where's the quality in systematic review? *Journal of Educational Policy* 20(4): 393-416
- Pawson, R (2006) *Evidence-based policy: A realist perspective*. London: SAGE
- Philpott, C (2017) Medical models for teachers' learning: asking for a second opinion. *Journal of Education for Teaching: International Research and Pedagogy* 43(1): 20-31
- Philpott, C and Poultney, V (2018) *Evidence-based teaching: A critical overview for enquiring teachers*. London: Critical Publishing
- Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised effect sizes. *Journal of Education Policy*, 32(4), 450-466
- Simpson, A (2018) Princesses are bigger than elephants: Effect size as a category error in evidence-based education. *British Educational Research Journal* 44(5)
- Thomas, G (2004) Introduction: evidence and practice. In G Thomas and R Pring (eds) *Evidence-based practice in education*. Maidenhead: Open University Press
- Thomas, G and Pring, R eds (2004) *Evidence-based practice in education*. Maidenhead: Open University Press
- Torgerson, C (2001) The need for randomised controlled trials in educational research. *British Journal of Educational Studies*, 49(3), 316-28
- Vygotsky, L (1925) *Consciousness as a problem in the psychology of behavior* (transl. N Veresov) <https://www.marxists.org/archive/vygotsky/works/1925/consciousness.htm>
- Worth, J, Sizmur, J, Walker, M, Bradshaw, S and Styles, B (2017) Teacher observation: Evaluation report and executive summary, November 2017. Education Endowment Foundation https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Teacher_Observation.pdf
- Wrigley, T (2018) The power of 'evidence': reliable science or a set of blunt tools? *British Educational Research Journal* 44(3)