BERJ British Educational Research Journal

BERA BRITISH EDUCATIONAL RESEARCH ASSOCIATION

*British Educational Research Journal*
*Vol. 44, No. 3, June 2018, pp. 359–376*

DOI: 10.1002/berj.3338

# The power of 'evidence': Reliable science or a set of blunt tools?

Terry Wrigley*
*Northumbria University, UK*

In response to the increasing emphasis on 'evidence-based teaching', this article examines the privileging of randomised controlled trials and their statistical synthesis (meta-analysis). It also pays particular attention to two third-level statistical syntheses: John Hattie's *Visible learning* project and the EEF's *Teaching and learning toolkit*. The article examines some of the technical shortcomings, philosophical implications and ideological effects of this approach to 'evidence', at all these three levels. At various points in the article, aspects of critical realism are referenced in order to highlight ontological and epistemological shortcomings of 'evidence-based teaching' and its implicit empiricism. Given the invocation of the medical field in this debate, it points to critiques within that field, including the need to pay attention to professional experience and clinical diagnosis in specific situations. Finally, it briefly locates the appeal to 'evidence' within a neoliberal policy framework.

**Keywords:** evidence; evidence-based teaching; EBM; randomised controlled trials; meta-analysis; empiricism; critical realism

## Introduction

The need for professionals to draw on evidence rather than political authority or custom and practice is not difficult to argue. However, like other feelgood words ('Intelligence', 'School Effectiveness', 'Leadership', 'Accountability', 'Standards'), it is important to interrogate the meanings they carry, and specifically within the current policy framework. Such keywords help produce ideological effects precisely because they appear beyond question, making it harder to investigate their inflection and deployment.

In the case of 'evidence'—more precisely, *evidence-based practice*—the ideological effect is reinforced by the cultural status of numbers in the modern era; numerical data are presented as objective, unmediated, unbiased and scientific carriers of facts (Poovey, 1998). In education in particular, under the sway of audit culture (Power, 1997), we have seen an escalation from assessment-based accountability to 'policy as numbers' and 'governing by numbers' (Ozga & Lingard, 2007) and now to demands for 'evidence-based teaching'.

What now stands proxy for a breadth of evidence is statistical averaging. This mathematical abstraction neglects the contribution of the practitioner's accumulated experience, a sense of the students' needs and wishes, and an understanding of social and cultural context. We see the attempted displacement of a rich array of research by the

*School of Health, Community and Education Studies, Northumbria University, Newcastle-upon-Tyne NE1 8ST, UK. E-mail: terrywrigley@gmail.com. Twitter: @terrywrigley1.

'gold standard' of randomised controlled trials (RCTs) and their statistical synthesis —a more appropriate term than 'meta-analysis', since it generally offers little by way of analysis. When 'evidence' is reduced to a mean effect size, the individual person or event is shut out, complexity is lost and values are erased.

This article aims to examine some of the technical shortcomings and ideological effects of demands for 'evidence-based teaching'. Firstly, given the argument that teachers should emulate evidence-based medicine (EBM), it is useful to recognise the contestation within that field. Secondly, the claim that evidence of this kind will make educational decision-making more 'scientific' is questioned. The article then focuses on the implications of methodological (and by implication epistemological and ontological) problems inherent in the privileging of randomised controlled trials and 'meta-analysis', as well as recent attempts at meta-meta-analysis such as Hattie's (2009) *Visible learning* project and the Education Endowment Fund's (EEF's) *Teaching and learning toolkit* (subsequently *Toolkit* for short). This critique then briefly situates the appeal to 'evidence' within the neoliberal policy framework.

At key points, aspects of critical realism are discussed in order to highlight ontological and epistemological shortcomings of 'evidence-based' methods, and the simplification and reductionism of their approach to causation.

## Evidence-based medicine

Given the recurrent calls for teachers to emulate doctors' strong use of evidence, it is important to understand that all is not so straightforward in that field either. Greenhalgh *et al.* (2014) argue that *real* evidence-based medicine:

- makes the ethical care of the patient its top priority
- demands individualised evidence in a format that clinicians and patients can understand
- is characterised by expert judgement rather than mechanical rule following
- shares decisions with patients through meaningful conversations
- builds on a strong clinician–patient relationship and the human aspects of care
- applies these principles at community level for evidence-based public health.

These experts contend that doctors need 'a more nuanced clinical expertise that embraces accumulated practical experience, tolerance of uncertainty, and the ability to apply practical and ethical judgement in a unique case'.

The problem of simplification has been well recognised in medicine, where, for good reasons, evidence-based practice is strongly established. Trish Greenhalgh, Professor of Primary Care Health Sciences, tells with some amusement her own story of ending up in hospital following a bicycle accident (Greenhalgh, 2015). She was immediately categorised as 'an elderly female who has had a fall', though she had been an amateur competitive cyclist and remained physically active. Although the evidence *in general* suggested one particular treatment, she was able to find more precise case studies to show why that would be inappropriate and have her treatment changed. Greenhalgh is clearly not suggesting we should dispense with evidence, but is calling for greater precision and for doctors to integrate it with well-developed clinical diagnostic skills.

A more fundamental accusation, in terms of research methodology, was made over 20 years ago when Feinstein (1995) described meta-analysis as 'statistical alchemy for the 21st Century'. He justified this charge by pointing to the requirement, in modern chemistry, to identify substances precisely and avoid working with impure mixtures. He accused meta-analysts of being unscientific by mixing together all manner of primary research studies to calculate an average effect size (see later sections for further discussion of evidence-based medicine).

## Making teaching 'scientific'

Although the concept of evidence-based teaching has been around for several decades, a significant intervention was made by Tom Bennett through his book *Teacher proof* (Bennett, 2013) and by founding ResearchED. Ostensibly these initiatives were intended to give new voice to teachers, but in effect serve to blinker them. Bennett's book begins soundly enough by pointing to the lack of evidence behind fads such as brain gym and damaging half-truths such as the visual–auditory–kinesthetic learning model, but then launches into a strident attack on all educational research that is not based on the 'gold standard' of RCTs. Rather than target the marketised provision of continuing professional development, he attacks academic university-based research. Bennett's stance was rapidly endorsed by schools minister Nick Gibb, who went so far as to claim the credit for the existence of ResearchED (Gibb, 2015), even though it purports to be a grassroots movement.

Bennett simplistically equates RCTs with science, forgetting that much scientific discovery has not arisen from experiments. Many scientific fields use few experiments (astronomy, meteorology, evolutionary theory—perhaps biology as a whole), and many discoveries and inventions did not arise from systematic procedures (e.g. penicillin, nylon, superconductivity) (Thomas, 2004). In the experimental sciences, close observation and theory play an essential part in articulating causality, which is not established by measuring regularity alone. Furthermore, there are crucial issues which distinguish the social from the natural world, including the matter of human agency.

We cannot simply write off the claims of quantitative research, including statistical methods. Radhika Gorur (2015) reminds us that numbers and statistics help produce 'calculable worlds'. They enable us to see a stretch of reality at a glance:

> What these entities lose in becoming detached from their contexts, they gain in becoming commensurate and combinable. (Gorur, 2011)

However, the converse is also true. As with all forms of symbolic representation—whether (to varying degrees) concepts, models or maps—abstraction highlights some details but loses others. Statistics developed as a way of seeing across an entire country. Indeed, *state* and *statistics* are related terms. Statistical procedures were developed to make territorial government manageable through a process of mapping chosen features. Even more so now that educational statistics aspires to a global reach, its measurement processes require a smoothing out of differences along with some mistranslation, a standardisation which approximates and distorts, and the use of invented categories and proxy indicators which are often misleading.

The process of measurement and calculation is not simply receptive, but can change the reality it purports to measure (Scott, 1998). We can see this in operation in English schools, as new undifferentiated categories soon appear as self-evident entities ('white British', 'FSMever', 'expected progress'), thus erasing the particular biographies, intentions and cultural assets of the student.

'What works' research also has a curricular effect. Our first question should be 'to what ends', but also 'in which situation' and 'for whom'? Aims in education are unsettled, contested and multi-layered. What might impact positively in terms of one aim could be harmful in terms of others. Appeals to 'evidence' depend on a shared understanding of the purposes of educational activity.

In significant ways, educational statistics has sought to straighten out the world in order to make it measurable. The school accountability system in England depends on assumptions of fairly regular and reliable linear progress—assumptions which recent reanalysis has undermined (see Reclaiming Schools, 2017 for a summary). However, this is only one aspect of the reductionism which the current version of 'evidence' entails. Another aspect is a lack of awareness of its own political location in a forcefield of disciplinary power (Foucault, 1977). Within a neoliberal policy framework of marketised competitive schools, the demand for 'evidence' becomes a further incursion on professionalism, part of the state's demand for greater efficiency in producing the next generation of human resources (Sears, 2003; Ball, 2013). As with much neoliberal ideology, it erases human history and culture through its appearance of value-free scientific neutrality (O'Neil, 2016).

## Level 1: Randomised controlled trials

Simplification is present even at the foundational level of specific pieces of empirical research, and key explanatory factors are systematically eclipsed. In one sense, this happens in laboratory experiments, which in a particular sense could also be regarded as reductionist. Steven Rose (2005: 73–97) argues that scientists make *tactical* use of simplification in constructing experiments but that they then have a responsibility for explaining and reconstructing the complexity of the real world. Problems arise when statisticians working in education or other social fields start to regard their simplified realities as a faithful mirror of the real world rather than an approximate sketch or topological map.

This certainly has to be guarded against in the field of medicine. The attempt to construct a clean experimental situation can introduce distortions which partly invalidate the findings. Richard Lehman, an experienced general practitioner, points out that real-life patients who come in with heart problems:

> have a median age of 76, equal gender mix and half of them have pretty good heart function and they invariably have other things wrong with them – what we call comorbidity. On the other hand, in so-called 'landmark trials' of heart failure drugs the median age of patients is 63, between 70 and 90 percent are male, and they are actually recruited for poor measures of heart function. In other words, they are younger but sicker, and comorbidity is an exclusion criterion. In other words, you're not allowed in the trial if you have anything else wrong with you. So of course the results from such randomised controlled trials cannot be applied directly to real patients. (Cited by Greenhalgh, 2016)

RCTs are an attempt to emulate the scientific rigour and neutrality of laboratory science in other fields, but there are multiple problems. In drugs trials, it is accepted practice to establish a control group which does not receive the treatment, to choose the experimental and control groups in ways which are free of human influence and, at best, to ensure 'double blinding' so that neither staff nor patient know whether individuals are taking the trialled drug or a placebo. This is clearly more difficult in education. Children are already allocated to classes and not easily individually reallocated. It is almost impossible to alter practice without the teacher or students noticing. Finally, there is no parallel to a *placebo*: should the control group experience the *absence* of the practice being trialled, or simply 'business as usual'?

This ambiguity concerning the control group can seriously distort attempts to calculate an 'effect size'. Imagine, for example, an RCT in using more open questions. Would the control group experience only closed questions, or would the teacher simply be asked to do as s/he normally does and not think too much about the type of question? These alternative possibilities could make a substantial difference to the effect size, as the former would probably increase the difference between the experiences of the two groups. As Pawson states:

> 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)

Several of these problems are illustrated by a recent project to evaluate *Fresh Start*, a product designed to remedy the reading difficulties of students in their first year at secondary school (EEF, 2015; Gorard *et al.*, 2016). The Executive Summary (EEF, 2015) asserts an effect size of +0.24SD between pre-test and post-test, roughly equivalent to 3 months' additional progress. The researchers acknowledge that there was a problem in allowing the schools to allocate pupils to control and treatment groups. However, those readers who dig deeper into the report will discover that there was such an imbalance between the groups that the mean *post-test* score of the treatment group was only slightly above the *pre-test* score of the control group.

To explore this further, the researcher then identifies matched subsets of both the control and treatment groups consisting of pupils with very low pre-test scores (i.e. roughly a third of pupils in the control group, and just over half of those in the treatment group). *Both* these subsets made an even larger gain than the *whole treatment group* did. In fact, both subsets are almost identical in terms of the mean pre-test score, the average gain and the mean post-test score. There is, in truth, no evidence of benefit from *Fresh Start* compared with the control group: the headlined 'three months additional progress' is simply a phantom of poor randomisation.

Further issues emerge from a close reading of the research report. Firstly, there is the question of agency:

> Participation in the Fresh Start intervention was at the instigation of the school leaders and cluster heads. They were already enthusiastic about the programme. (p. 5)

And later:

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

In other words, any apparent benefit from using the Fresh Start system may well be largely a product of classroom ethos and emotion.

Secondly, there is no attempt in this research to investigate or diagnose the reasons why particular pupils are having difficulties learning to read. This remains within the black box of aggregate statistical data. Thus we are no nearer an understanding of why some remedial interventions might be more successful than others. In other words, in critical realist terms, there is no attempt to find the 'mechanism'. This blind empiricism is deprofessionalising in its disconnect from pedagogical reasoning.

Finally, there is considerable ambiguity surrounding the control condition. We do not learn whether teachers and teaching assistants in the control group had any access to training comparable to that of the treatment group, whether they also taught small classes, or what 'business as usual' actually involved.

The doubts being raised here are not only technical, but have political implications. Firstly, it is clear that the supposed rigour of RCTs (which the EEF, on the government's behalf, makes a condition of research funding unless this proves impossible) has been evaded here. Secondly, we should note that the teaching method exemplified in *Fresh Start*, namely synthetic phonics, is strongly promoted by the schools minister Nick Gibb and vigorously sponsored and protected by various policy measures, including the statutory Phonics Check. This raises the possibility that the decision to headline 'three months additional progress' rather than 'no demonstrable benefit' could have been imposed on the research team through direct or tacit political pressure exercised via the funding agency.

[This augurs badly for the objectivity of another EEF (2017) project, also concerning synthetic phonics: £1m is to be shared between a Northern Ireland university and Ruth Miskin Training to evaluate her own company's reading schemes. Although two US-based external evaluators are listed, there is an obvious danger of distortion occurring before their involvement, or which might go unnoticed.]

### *Open systems and the human factor*

These basic technical issues of randomisation, agency (teacher enthusiasm) and how the control groups are taught are only the start of the difficulties. As Pawson (2006: 18) argues, social situations are 'open systems', the product of multiple components and forces:

> A ceaselessly changing complexity is the norm in social life, and this is the open system predicament.

Behaviours are shaped by historical forces and cultural norms, the institutions we inhabit and the choices of individuals. Moreover, 'even the research act itself is transformative; social research always has the tendency to disturb what it is trying to describe'. It is hardly surprising, therefore, that it is difficult to isolate a single factor and stabilise the rest.

Pawson emphasises that whereas drugs trials try to eliminate the human factor because 'human volition is seen as a contaminator' (Pawson, 2006: 27), social change is brought about *through* the human agent:

> Social programmes... offer resources (material, social, cognitive) to subjects, and whether they work depends on the reasoning of these individuals. Subjects may seek out programmes (or not), volunteer for them (or not), find meaning in them (or not), develop positive feelings about them (or not), learn lessons from them (or not), apply the lessons (or not), talk to others about them (or not). It is within this interpretative process - or mechanism - that the causal powers of programmes reside. (Pawson, 2006: 45)

This problem of agency is well illustrated by a recent attempt to evaluate project-based learning within the required norms of EEF funding (EEF, 2016). This trial involved 12 intervention and 12 control schools, altogether around 4,000 Year 7 students, occupying 25–50% of the timetable for almost a year. The high dropout rate (nearly half the intervention schools) suggests a problem in convincing teachers and also perhaps students. This raises the question of whether pedagogies requiring a strong professional commitment can be evaluated through such a trial, especially when they break from the tightly controlled pedagogies which have become the norm in a high-stakes accountability regime. More broadly, we are faced with a paradox which puts the entire RCT methodology in doubt as far as education is concerned: human volition is both *necessary* and a *contaminator*. The question of agency and pedagogical intention is inescapable, and will be discussed later through the lens of critical realism.

### *Regularity and causality: Hume's dilemma*

In many social situations, the extent of variation from a perceived pattern can be as interesting as the regularity, but this is inadequately captured by summative indicators such as standard deviation, range or effect size. Hubert and Wainer (2013: 119) insist that:

> In any reasoning based on the presence or absence of a correlation between two variables, it is imperative that graphical mechanisms be used in the form of scatterplots. One might go so far to say that if only the value of $r_{XY}$ is provided and nothing else, we have a prima facie case for statistical malpractice.

Beyond the technical aspects, we need to consider the philosophical implications of RCTs and specifically the relationship between regularity and causality. This goes beyond the usual warning that 'correlation does not imply causation' as there may be other 'lurking' or third variables at work driving both X and Y. David Hume argued that however many times one billard ball hits another and the second ball moves, this cannot prove causality. This is, implicitly, the underlying stance of much educational statistics, which contents itself with observing regularities without seeking causal explanation.

This Humean empiricist stance is opposed by both 'scientific realists' and 'critical realists', for the reasons presented clearly by Roy Nash:

> Scientific realism rejects the standpoint of 'positivist' science, with its Humean negation of causality, its construction of models in terms of laws with no necessary reference to mechanism, and its indifference to the essence and substance of things, as inadequate to a satisfactory explanation of physical and social events and processes. (Nash, 2002: 398)

Nash points to other shortcomings in statistical methodologies, including a failure to recognise that some entities and activities have only 'weakly quantitative properties' (p. 401) and the belief that elaborate statistical procedures are sufficient to identify causal models without an investigation of the qualities of the entities and their relationships. Consequently, Nash argues for the complementary investigation of 'numbers and narratives'.

Bhaskar (1978) and other *critical realists* overcome the blockage caused by Hume's sceptical empiricism by distinguishing (1) the real, (2) the actual and (3) the phenomenal. Critical realism provides a means of understanding causality beyond a simple repetition or regularity: we may not always experience or observe (3) what actually happens (2), and furthermore the underlying forces (1) may fail to actualise (2) in open systems. Thus, the purpose of scientific experiments is to make these forces (1) visible (3). This is not, of course, an argument for occult or mysterious forces, but a call for 'in-depth' realism.

Sayer summarises the implications as follows:

> The conventional impulse to prove causation by gathering data on regularities, repeated occurrences, is therefore misguided: at best these might suggest where to look for candidates for causal mechanisms. What causes something to happen has nothing to do with the number of times we observe it happening. (Sayer, 2000: 14)

Using the simple example of gunpowder, causal mechanisms are partly inherent in the substance:

> gunpowder has the tendency to go off with a bang because of what it is... The chemical composition generates the capacity to explode. (Pawson, 2006: 23)

But causality is also located in the context, so that the strength of an explosion, indeed whether or not it happens, depends on factors such as moisture, temperature and pressure. Bhaskar (1998) goes further than this to argue that human aims, beliefs and intentions should also be understood as causal (Aristotle's 'final causes'). The critical realist demand for more complex models of causation, beyond the methodological simplicity of 'evidence-based teaching', will be discussed further in later sections.

None of these arguments, in finality, invalidate RCTs as a method but should lead to a more guarded understanding of their role. In education, this certainly undermines Bennett's claim that other forms of research are worthless. Firstly, we should not underestimate the importance of attentive observation in natural and social sciences:

> Although the other physical and biological sciences have achieved great advances by supplementing observation with controlled experimentation, qualitative observation plays a critical and foundation role in every scientific area in the formation of theory and hypotheses, the design of research projects, and the exploration of new frontiers. (Lingenfelter, 2016: 114)

Lingenfelter also reminds us that, in evaluations of teaching, qualitative methods enable us to follow the 'perspectives and observations of multiple participants and observers'.

**Level 2: Meta-analysis**

There is an old joke about the man found lying with his head in the oven and his feet in the fridge. A statistician comes along and declares that *on average* his temperature is perfectly normal.

The process by which much of the complexity of change in real situations is smoothed out during RCTs is compounded when these are bundled together. Rather than engaging in a critical discussion of available research, the current fashion, privileged by major fundholders such as the EEF, is for 'meta-analysis'.

The decisions about which original research studies should be included are based on *technical* rather than substantive criteria. One consequence is that quite dissimilar studies are thrown together and an aggregate mean of effect sizes calculated. Although some tolerance is acceptable in meta-analysis, since no two research studies are exactly alike, serious problems can arise from aggregating and averaging studies using different definitions of an issue, and based on different curriculum areas, ages and attainment levels of students, types of school, education systems, and so on.

The problem is commonly referred to as mixing *apples and oranges*, and was identified many years ago in the medical literature. In medicine, Feinstein complains that vital information is omitted (e.g. severity of illness) and that differences are lost when data is merged. Even where findings diverge strongly, or go in opposite directions (some studies showing positive and some showing negative effects), the disparate effect sizes are often lumped together and averaged. Feinstein (1995) insists that we should 'stop believing the often stated dogma that "randomization prevents bias"': instead 'Important inconsistencies are ignored and buried in the statistical slurry'.

Indeed, Gene Glass, who originated the idea of meta-analysis, issued this sharp warning about heterogeneity:

> 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)

Within education, Robert Coe once issued a similar warning, though this is now *systematically* ignored in the Toolkit produced by his organisation for the EEF. It is worth reading this at length:

> One final caveat should be made here about the danger of combining incommensurable results. Given two (or more) numbers, one can always calculate an average. However, if they are effect sizes from experiments that differ significantly in terms of the outcome measures used, then the result may be totally meaningless. . .
>
> In comparing (or combining) effect sizes, one should therefore consider carefully whether they relate to the same outcomes. . . One should also consider whether those outcome measures are derived from the same (or sufficiently similar) instruments and the same (or sufficiently similar) populations. . . It is also important to compare only like with like in terms of the treatments used to create the differences being measured. In the education literature, the same name is often given to interventions that are actually very different. It could also be that. . . the actual implementation differed, or that the same treatment may have had different levels of intensity in different studies. In any of these cases, it makes *no sense to average out their effects.* (Coe, 2002, my italics)

A good example in the Toolkit is the blanket category of 'feedback', used to gather together many different forms of teacher intervention, advice and formative assessment. From a plethora of studies with divergent effect sizes, some of them negative, an aggregate mean effect size is calculated and the conclusion drawn that 'feedback' is the most effective way to improve attainment. Since feedback is inevitably present in some way in any pedagogical interaction, it would be more illuminating to examine reasons for the differences.

Hattie comes close to a similar problem in his clear preference for 'direct instruction' over enquiry-type methods. However, provided one reads his words closely, he is actually referring to a specific model of 'direct instruction' (from Adams & Engelmann, 1996) rather than general notions of didactic teaching, teaching from the front or rote learning. This model involves not only clear presentation, but a sequence of learner engagement, modelling, guided practice, monitoring and independent practice/transfer (Hattie, 2009: 205–206). Although he states (208–212) that inquiry methods and problem-based learning are less efficient for learning facts and concepts, he agrees that they are better for longer-term recall, understanding the principles that link concepts together, engaging students, applying knowledge, solving problems, critical thinking and scientific process. Unfortunately, the dials which decorate these pages can be 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.

This points to a major deficit in meta-analysis: that of theoretical explanation. Pawson complains that statistical research in social fields is often undertheorised—unlike in medicine:

> Medical treatments... are the embodiment of years of theory-testing. They are already scientific inquiry incarnate before the first Phase III RCT is even designed. By this stage, medical science knows pretty well how a treatment works and it entrusts to the RCT a slightly different question about how well it works in a particular manifestation. Whole episodes of pure science are played out, and their lessons digested, before the applied science kicks in. (Pawson, 2006: 47)

Pawson argues, therefore, that systematic reviews of research must seriously work on developing an:

> understanding of how interventions work. Theory-testing remains essential in each evaluation and each review. We need to persist in asking how an intervention works in order to figure out how well it works. The better meta-question is an explanatory one. (Pawson, 2006: 47)

## Level 3: Meta-meta-analysis

It is hardly surprising that *Visible learning* (Hattie, 2009) and related books are international best-sellers. The prospect of having at your fingertips a summary of all you ever need to know is seductive for busy teachers, school leaders and administrators alike. The graphic device of a dial resembling a car's speedometer adds to this seductive effect: you can see the effectiveness *at a glance*. The project is a synopsis of 800 meta-analyses based on over 50,000 separate research studies. Apart from the sheer

hubris of the claim to have intelligently analysed such a broad field, we need to be aware of some specific problems:

- The source studies are overwhelmingly from the USA, and up to 50 years old.
- Many use narrow outcome measures which do not reflect important educational aims (e.g. reading aloud single words as proxies for reading, or basic arithmetic questions for maths).
- There is a serious underrepresentation of curriculum areas beyond basic literacy and numeracy.

All the warnings in the previous section of this article apply here too—averaging does not wipe them out—but there are more. Firstly, there is confusion around *effect size*, including Hattie's notion of a 'hinge point' of 0.4. Part of his logic is that the average annual improvement by students is an effect size of 0.2 to 0.4, hence his argument that we should discard any intervention with a smaller effect size. However, this premise has been sharply questioned by other statisticians:

(i) No account is taken of the duration of each intervention, which could vary from a few weeks to a year or more. (Brown, 2013)
(ii) Diverse outcomes are jumbled together, including literacy, numeracy, other specific curriculum areas and psychological gains. (Brown, 2013)
(iii) No allowance is made for the tendency of average effect sizes to reduce dramatically with the children's age from 5 to 9 years. (Orange, 2014a, b)
(iv) The calculations vary in breadth, from very specific to broad categories, and 'causal factors' such as 'home', 'personality', 'parental involvement', 'happiness' are juxtaposed with specific teaching methods. (Higgins & Simpson, 2011)
(v) Sometimes Hattie uses 'effect size' to mean 'as compared to a control group' and at other times to mean 'as compared to the same students before the study started'. (Literacy in Leafstrewn, 2012)

Hattie is at his best when he engages in a more reflective and precise analysis of specific research studies. As an illustration, Hattie and Yates (2014: 72–83) includes a table which is entirely misleading, since it refers in very general terms to 'effect sizes on achievement' without specifying *what* is achieved; nevertheless, once we look beyond this, for example at the detailed reports of research into history teaching (critical understanding of primary sources) and biology teaching (explicit consideration of alternative explanations, understanding experimental limitations), we find some illuminating analysis.

The technical problems of effect size in meta-analysis are further exposed by Adrian Simpson (2017), this time with specific reference to the Toolkit. The Toolkit uses the visual device of a league table, which ranks interventions by 'additional months of progress' derived from effect sizes. (There is clearly some difficulty here, since implementing all of them simultaneously would add more than 8 years.) It describes nearly 30 kinds of intervention, which range across teaching methods to school organisation to supplementary activities. Each of these is treated as a meta-analysis in its own right, and the average effect size calculated. Altogether, the Toolkit is intended as a kind of meta-meta-analysis designed to support headteachers' decisions about how to spend additional budgets to reduce the poverty-related attainment gap.

Simpson argues that the differences in effect size are as likely to be a function of technical difficulties, rather than showing a real difference of impact. He explains three major sources of inaccuracy:

(i) *Comparison groups*—the lack of clarity about the control group's activity (whether 'business as usual' or a zero condition) affects the measured 'effect size'.
(ii) *Range restriction*—research based on a limited population (e.g. 11-year-old boys with reading difficulties) will tend to show a larger effect size. (Because the range, shown as SD, is the denominator in the formula, a reduced range automatically magnifies the result.)
(iii) *Measure design*—trials which use outcome measures closely related to the nature of the intervention will show a larger effect size than where the outcome measure is more general in nature.

Simpson pursues this explanation in detail, with numerous examples from the Toolkit, many of which could have serious practical consequences in terms of management decisions. For example, studies which conclude that computer use is particularly beneficial for pupils with special educational needs might suffer from inflated effect sizes due to the smaller range. Some studies of maths teaching use outcome tests which are closely related to the intervention (for example, an aspect of algebra), whereas others use a broad-brush standardised test.

One of the lowest-rated categories in the Toolkit is 'teaching assistants'. Different social contexts, age groups and pupil needs are merged, but the most significant source research was led by Peter Blatchford, who chose to speak back. His research, in fact, pointed to classroom assistants working in conditions where no time was given for guidance from the teacher or for evaluation afterwards. It complained of classroom assistants always being assigned to lower attainers, thus depriving these children of help from a qualified teacher. Blatchford was *not* suggesting that classroom assistants are ineffective, but pointing to ways in which they could bring *greater benefit*. We should also recognise that classroom assistants serve a range of purposes, not all of which are measured through attainment. Clearly, placing classroom assistants near the bottom of the Toolkit's league table, with a label of 'low impact for high cost', could result in schools and academy chains terminating their employment, especially in times of budget cuts.

Given these problems, it is only by chance if aggregation brings sound results. Whilst some conclusions may be tactically appealing, for example the low ratings for government-approved practices such as performance pay and streaming/setting, it can be extremely misleading. Admittedly the Toolkit's authors urge caution:

> The evidence it contains is a supplement to rather than a substitute for professional judgement: it provides no guaranteed solutions or quick fixes... We think that average impact elsewhere will be useful to schools in making a good 'bet' on what might be valuable, or may strike a note of caution when trying out something which has not worked so well in the past. (Higgins *et al.*, 2012)

However, many busy teachers and heads will inevitably take the league table at face value and remain unaware of its many problems.

It is not exactly the case that the Toolkit's authors believe the mean effect sizes are sufficient in themselves. Indeed, the summary is now supplemented by expanded

pieces of advice to teachers and school management. Ironically, however, this depends as much on personal experience, good judgement and instinct as on its approved model of research.

In summary:

(i) The league table format systematically encourages aggregation of dissimilar studies (apples and oranges).
(ii) Many interventions depend on context, and will not work as well as in the trial situation.
(iii) There is some misrepresentation of research.
(iv) The precise nature of interventions is generally invisible, as is also the teaching experienced by the 'control group'.
(v) Because the focus is exclusively on attainment, and sometimes defined by narrow outcome measures, wider aims of education are eclipsed.
(vi) Finally, much of the primary research is not based on pupils similar to those whom the Toolkit is supposed to help (i.e. children suffering from *poverty-related disadvantage*).

## The dangers of simplification: Some implications of critical realism

There are many fruitful ways in which researchers can collaborate with educators, and for schools to draw on research. However, the recent 'evidence' cult has resulted in various forms of philosophical as well as technical simplification.

As Pawson forcefully points out, the problems do not disappear as one moves to second (meta-) and third (meta-meta-) levels of aggregation:

> At every stage of the meta-analytic review, simplifications are made. Hypotheses are abridged, studies are dropped, programme details are filtered out, contextual information is eliminated, selected findings are utilized, averages are taken, estimates are made... In this purgative process the very features that explain how interventions work are eliminated from the reckoning. Complex programmes are cast as simple treatments. The way in which stakeholders think and change their thinking under an intervention is expunged. (Pawson, 2006: 42–43)

Rather than becoming more powerful and informative in terms of causal explanations, it gets hollowed out. In terms of the earlier reference to critical realism and Hume's dilemma, there is insufficient attempt, within this paradigm, to dig down in search of *causes* (the 'deep real'), so the explanations remain at the level of *regularities* (the actual).

Despite the need to study carefully the irregularities in data which might reflect multiple causal factors operating in non-linear ways (discussed above), the emphasis remains, in practice, on the generality of 'what works' rather than *where* it might work, *for whom* and *under what conditions*. The sloganistic 'what works' reflects a neoliberal demand to extract maximum efficiency from education, while marginalising the qualitative and political dimensions of human formation (Ball, 2013).

The complexity of projects and situations is obscured through statistical meta-analysis, which becomes overstretched when operating beyond fairly straightforward linear relationships and closed systems. As Biesta (2010: 496) explains:

> 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.

This connects strongly with the ontologogical position of critical realism. Even in the natural world, closed systems are rare and have to be brought about artificially through the construction of experiments. This is even more so in the fields investigated by social sciences, where the capacity of human beings to make decisions, or even inadvertently to act upon beliefs, makes predictable regularity almost unobtainable. As Sayer (2000: 15) explains with an example from economics, 'the same causal power can produce different outcomes, according to how the conditions for closure are broken: for example, economic competition can prompt firms to restructure and innovate or to close'.

Sayer adds that regularities in social systems 'are usually the product of deliberate efforts to produce them, through devices such as disciplinary regimes, for example... machine-pacing of work' (Sayer, 2000: 15). This itself should raise the question of what kind of educational aims such disciplinary regimes might fulfil. Education has to be regarded as an open system, for pedagogical and ethical reasons, despite the desperate attempts of neoliberal policymakers to make it utterly measurable and predictable in the interest of maximising the production of future human capital.

In terms of providing guidance to practitioners and policymakers, we should note Pawson's proposals for a 'realist synthesis' of research (2006: 78–94), which aims to develop convincing theories of causation. Rather than simplify the original research and turn out an average effect size, Pawson advocates an enriched understanding of the 'subjects' of each intervention, the causal theories proposed by the original researchers, the quality of outcomes and the adequacy of measures, processes and blockages. Chris Brown and colleagues examine diverse examples of interaction and participation involving researchers, teachers and policymakers (Brown, 2014, 2015). Similarly, Lingenfelter (2016: 118 ff) provides a useful summary of Bryk's development of *networked improvement communities* and their collaborative development of knowledge, structures and action (Bryk *et al.*, 2010).

These various approaches to educational change cry out for something approximating to a model of stratification for school life, learning and governance. Whereas the Toolkit and 'Visible learning' imply a miscellany of separate arrangements or interventions, each having a discrete 'effect size', the field of school-based education can be viewed in terms of different levels of activity or situation which act on and through one another: governance, assessment, school ethos, regulations, pedagogy, various kinds of relationship, 'leadership', professional development, and so on. What is the relationship, then, between national regulations, school structure and culture, classroom pedagogies, students' cultural assets? The force which particular factors from the different 'strata' exert is not simply additive, since they can accelerate or negate one another.

Dealing with all this may not be as straightforward as calculating meta-analytic mean effect sizes, or have the rhetorical power of 'months of additional progress' rankings or dashboard dials, but it could result in reliable practical knowledge while at the same time strengthening teachers' capacity to reflect on their practice. This richer and more theorised kind of research synthesis, and its collaborative interpretation, may not provide *instantly visible* answers or enable policies to be selected at a glance, but it will benefit education far more than 'a good *bet* on what might be valuable' (Higgins *et al.*, 2012) based on a Toolkit filled with blunt tools and misleading summative data.

We should also question the notion of 'intervention' in educational settings, and ask whether this implies a transmission model whereby something is 'done to' the learner, or perhaps the 'banking' model criticised by Freire. We could contrast this with various cultural–historical (CHAT) pedagogies based on the engagement of learners with reality mediated by cultural tools, social structures and learning communities (e.g. Engeström, 1999). Practices based on cultural–historical principles challenge passivity and mechanical models of school change, and place meaning-making back at the heart of educational research. They remind us of the agency, and potential for resistance, of the learner as well as the teacher. They remind us that cognitive development depends on the ability of the teacher to help learners engage with words and other forms of representation—'cultural tools'—in order to shed a different light on phenomena, to dig below surface appearances, to grapple with underlying forces and structures. They enable us to work with rich concepts of 'school culture' as a conjuncture of objects, rituals, interactions and habits, which carry and convey meaning to participants with agency, and which build an environment in which powerful learning can *emerge*.

Critical realism—to recapitulate and extend earlier points—provides an ontology and epistemology whereby *underlying* forces interact in rather unpredictable ways and might or might not actualise. These forces and structures sit in various *strata* of reality, and through their energy, interactions and engagements with the environment, new possibilities of *emergence* arise: unlike the multiple 'effect sizes' of meta-meta-analysis, the sum can be more than the parts and qualitative change can arise (summarised by Banfield, 2016: 98). Causation is complex, involving the properties of matter, structures, patterns of cause and effect and, crucially, human beliefs, desires and intentions (Banfield, 2016: 100 ff, drawing on Aristotle). This is a worldview where the past matters, whether as habitus or culture or conscious understanding. The concept of emergence and the importance of human intention point to the *emancipatory* capacity of education, rather than a focus on efficient 'delivery' of fixed knowledge. Critical realism involves an ontology where meaning and meaning-making have a central role and are inter-imbricated with structures, actions and environments. All this offers far more powerful models than the 'flat-world' empiricism of evidence-based teaching.

The insistence that research must centre on 'what works' obscures questions of educational purpose: *what* is it that we want to work, and *why*? It is precisely these questions which neoliberalism would prefer to avoid, so that everything becomes a technical question and value questions are marginalised. The key point is that education is about more than the acquisition of knowledge and skills: it is about

understanding and wisdom, living well and learning how we can live together as human beings and have a sustainable future on planet earth. This is why research too needs to be broadly conceived.

The privileging of supposedly rigorous research about the technical efficiency of teaching techniques is reductionist in its operation and implications: it simplifies the complexity of social organisation and relationships, it reduces young people to receptacles for fragments of curricular knowledge, it narrows the aims of education and reduces curricular range. It deprofessionalises teachers by blocking genuine interaction and participation in research, all the while claiming to give teachers voice and to upgrade the profession. It homogenises learners, teachers and schools in the interest of 'effectiveness' and making schools manageable.

Numbers in themselves are not the problem. As Espeland and Stevens (2008: 432) point out:

> Measurement can help us see complicated things in ways that make it possible to intervene in them productively (consider measures of global warming); but measurement also can narrow our appraisal of value and relevance to what can be measured easily, at the expense of other ways of knowing.

The problem comes from an inflated and generalised role for statistical studies, a lack of awareness and self-awareness, and the omissions and linearities that arise in order to create an aura of science, order and regularity. The attempt to make learning *visible* (as Hattie puts it) eclipses older understandings of education as Bildung and pedagogy (both words carrying the sense of human formation). It serves to make *invisible* the deep aims of education, in terms of what kind of human beings we are forming and what kind of future we hope for.

In the face of attempts to narrow down educational research to clumsy calculations of efficiency, we need to argue the importance of a wide methodological spectrum. Pawson (2006: 50) argues not only for the importance of qualitative evidence in pursuing *how* teaching works, but also that 'the evidence base should include data procured by comparative research, historical research, discourse analysis, legislative inquiry, action research, emancipatory research, and so on'. Diverse forms of research are needed in order to answer questions about causality, human agency, social contexts, interactions and educational purpose, without which it is vacuous to speak of 'effectiveness'. There is no space here to discuss the role of philosophy, educational sociology, ethnography, critical policy studies, case studies, classroom observation; but all of these are important if we are to help return power to teachers, parents and young people, and see beyond the shallow functionalism of a 'dictatorship of no alternatives' (Unger, 2005).

## References

Adams, G. & Engelmann, S. (1996) *Research on direct instruction: 20 years beyond DISTAR* (Seattle, WA, Educational Achievement Systems).

Ball, S. (2013) *The education debate* (2nd edn) (Bristol, Policy Press).

Banfield, G. (2016) *Critical realism for Marxist sociology of education* (London, Routledge).

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, 491–503.
Brown, C. (2014) *Making evidence matter: A new perspective for evidence-informed policy making in education* (London, IOEPress).
Brown, C. (Ed.) (2015) *Leading the use of research and evidence in schools* (London, IOEPress).
Brown, N. (2013, August 5) *Book review: Visible learning*. *Academic Computing blog*. Available online at: academiccomputing.wordpress.com/2013/08/05/book-review-visible-learning/ (Accessed 18 March 2018).
Bryk, A., Gomez, L. & Grunow, A. (2010) *Getting ideas into action: Building networked improvement communities in education*. Available online at: www.carnegiefoundation.org/spotlight/webinar-bryk-gomez-building-networked-improvement-communities-in-education (Accessed 18 March 2018).
Coe, R. (2002) It's the effect size, stupid: What effect size is and why it is important, paper presented at the *British Educational Research Association Conference*, Leeds, 12–14 September.
EEF (2015) *Fresh Start: Evaluation report and executive summary*. Available online at: v1.educationendowmentfoundation.org.uk/uploads/pdf/Fresh_Start_(Final).pdf (Accessed 18 March 2018).
EEF (2016) *Project-based learning*. Available online at: educationendowmentfoundation.org.uk/projects-and-evaluation/projects/project-based-learning/ (Accessed 18 March 2018).
EEF (2017) *Read Write Inc. Phonics and Fresh Start*. Available online at: educationendowmentfoundation.org.uk/projects-and-evaluation/projects/read-write-inc-and-fresh-start/ (Accessed 18 March 2018).
Engeström, Y. (1999) Activity theory and individual and social transformation, in: Y. Engeström, R. Miettinen & R.-L. Punamäki (Eds) *Perspectives on activity theory* (Cambridge, Cambridge University Press).
Espeland, W. & Stevens, M. (2008) A sociology of quantification, *European Journal of Sociology*, 49 (3), 401–436.
Feinstein, A. (1995) Meta-analysis: Statistical alchemy for the 21st Century, *Journal of Clinical Epidemiology*, 48(1), 71–79.
Foucault, M. (1977) *Discipline and punish* (New York, Pantheon).
Gibb, N. (2015) *The importance of the teaching profession*. Keynote speech at ResearchED Conference, 5 September. Available online at: www.gov.uk/government/speeches/nick-gibb-the-importance-of-the-teaching-profession (Accessed 18 March 2018).
Gorard, S., Siddiqui, N. & See, B. H. (2016) An evaluation of Fresh Start as a catch-up intervention: A trial conducted by teachers, *Educational Studies*, 42(1), 98–113.
Gorur, R. (2011) ANT on the PISA trail: Following the statistical pursuit of certainty, *Educational Philosophy and Theory*, 43(S1), 76–93.
Gorur, R. (2015) Producing calculable worlds: Education at a glance, *Discourse: Studies in the Cultural Politics of Education*, 36(4), 578–595.
Greenhalgh, T. (2015) *Real vs rubbish EBM: What is the state of evidence-based medicine* (Oxford, CEBM).
Greenhalgh, T. (2016) Evidence-based medicine: A model to follow? (or not. . .), PowerPoint prepared for NUT/Rethinking Schools Seminar, *Teaching by Numbers: Accountability Data and 'Evidence Based Practice'*, 13 January.
Greenhalgh, T., Howick, J. & Maskrey, N. (2014) Evidence based medicine: A movement in crisis? *BMJ*, 348:g3725 (13 June).
Hattie, J. (2009) *Visible learning: A synthesis of over 800 meta-analyses relating to achievement* (London, Routledge).
Hattie, J. & Yates, G. (2014) *Visible learning and the science of how we learn* (London, Routledge).
Higgins, S. & Simpson, A. (2011) Visible learning: A synthesis of over 800 meta analyses relating to achievement. By John A. C. Hattie, *British Journal of Educational Studies*, 59(2), 197–201.

Higgins, S., Kokotsaki, D. & Coe, R. (2012) *The teaching and learning toolkit*. Available online at: v1.educationendowmentfoundation.org.uk/uploads/pdf/Teaching_and_Learning_Toolkit_(July_12).pdf (Accessed 18 March 2018).

Hubert, L. & Wainer, H. (2013) *A statistical guide for the ethically perplexed* (London, CRC Press).

Lingenfelter, P. (2016) *'Proof', policy and practice: Understanding the role of evidence in improving education* (Stylus, VA, Sterling).

Literacy in Leafstrewn (2012, December 20) *Can we trust educational research ('Visible Learning': Problems with the evidence)*. Available online at: literacyinleafstrewn.blogspot.co.uk/2012/12/can-we-trust-educational-research_20.html (Accessed 18 March 2018).

Nash, R. (2002) Numbers and narratives: Further reflections in the sociology of education, *British Journal of Sociology of Education*, 23(3), 397–412.

O'Neil, C. (2016) *Weapons of math destruction: How big data increases inequality and threatens democracy* (New York, Penguin).

Orange, O. (2014a, August 20) *The age effect which means the 'effect size' is useless*. *Ollieorange2 blog*. Available online at: ollieorange2.wordpress.com/2014/08/20/visible-learning-6-age-and-the-effect-size/ (Accessed 18 March 2018).

Orange, O. (2014b, September 24) *John Hattie admits that half of the statistics in Visible Learning are wrong (part 2)*. Ollieorange2 blog. Available online at: ollieorange2.wordpress.com/2014/09/24/half-of-the-statistics-in-visible-learning-are-wrong-part-2/ (Accessed 18 March 2018).

Ozga, J. & Lingard, B. (2007) Globalisation, education policy and politics, in: B. Lingard & J. Ozga (Eds) *The RoutledgeFalmer reader in education policy and politics* (London, Routledge).

Pawson, R. (2006) *Evidence-based policy: A realist perspective* (London, Sage).

Poovey, M. (1998) *A history of the modern fact: Problems of knowledge in the sciences of wealth and society* (Chicago, IL, University of Chicago Press).

Power, M. (1997) *The audit society: Rituals of verification* (Oxford, Oxford University Press).

Reclaiming Schools (2017) The illusions of measuring linear progress, in: *Beyond the exam factory: Alternatives to high-stakes testing* (Northampton, More Than A Score).

Robinson, D. (2004) An interview with Gene V Glass, *Educational Researcher*, 33(3), 26–30.

Rose, S. (2005) *Lifelines: Life beyond the gene* (2nd edn) (London, Vintage).

Sayer, A. (2000) *Realism and social science* (London, Sage).

Scott, J. (1998) *Seeing like a state: How certain schemes to improve the human condition have failed* (New Haven, CT, Yale University Press).

Sears, A. (2003) *Retooling the mind factory: Education in a lean state* (Aurora, Ont., Garamond).

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

Thomas, G. (2004) Introduction: Evidence and practice, in: G. Thomas & R. Pring (Eds) *Evidence-based practice in education* (Maidenhead, Open University Press).

Unger, R. (2005) *What should the Left propose?* (London, Verso).