Evidence-based teaching: a simple view of 'science'

Terry Wrigley and Sean McCusker, Northumbria University

Abstract

This paper examines the insistent claims by advocates of Evidence-Based Teaching, that it is a rigorous scientific approach. The paper questions the view that randomised controlled trials and meta-analyses are the only truly scientific methods in educational research. It suggests these claims are often based on a rhetorical appeal which relies on too simple a notion of 'science'. Exploring the tacit assumptions behind 'evidence-based teaching', the paper identifies an empiricist and reductionist philosophy of science, and a failure to recognise the complexity of education and pedagogy. Following a discussion of large-scale syntheses of evidence (Hattie's Visible Learning; the Education Endowment Foundation's Teaching and Learning Toolkit), it examines in detail one strand of the latter concerning Sports Participation, which is used to illustrate flaws in procedures and the failure to take seriously the need for causal explanations.

Introduction

Increasingly in recent years, the field of Education has seen strong and indeed dogmatic insistence on 'scientific' approaches to evidencing 'what works'. It is the aim of this paper to discuss these claims and the posture of its advocates, firstly in general terms and then through a detailed examination of how evidence is assembled in one particular study. It will argue that the appeal to 'science' as the model for valid educational research too often depends on a superficial vernacular notion of science.

A hierarchy of approved methodologies is proposed with Randomised Controlled Trials (RCTs) set up as the 'gold standard', their conclusions pulled together through a statistical synthesis of effect sizes (misleadingly known as 'meta-*analysis'*). Taking this one stage further to what we might call 'meta-meta-analysis', John Hattie (2009) has achieved international recognition for his attempt to synthesise over 800 meta-analyses involving 50,000 research studies. Actions and interventions are then compared on the basis of comparative effect sizes. More modestly, the Education Endowment Foundation (EEF) has produced a Teaching and Learning Toolkit (Higgins et al. 2014), a comparison of 35 different forms of action to evaluate their relative effectiveness in assisting pupils disadvantaged by poverty. In this case, the mean effect sizes are translated into months of additional progress, ranging from +8 to -4 months. The EEF also commissions many new RCTs.

The use of RCTs and meta-analysis is sanctified by appeals to Science in general, including more specific demands that teachers should learn from health professionals' adoption of Evidence-Based Medicine (EBM). Polemical appeals to 'scientific' rigour are circulating in various media, genre and spheres of influence. These draw on long established binaries which distinguish science as objective knowledge from subjective and even superstitious beliefs (Lakoff and Johnson 2003:187seq). Subjectivity is regarded as self-indulgent, exaggerating the importance of the individual; it uses poetic or rhetorical language which lacks precision, and privileges aesthetic, moral and spiritual sensibilities. This binary is at work in the 'scientific' claim used to give a positive intonation to approved forms of educational research (RCTs, meta-analysis). The other side of the binary, for example the discredited pre-scientific practices of alchemy or astrology, is used symbolically to designate and discredit other forms of research. Thus, some of the most strident advocates have attempted to rubbish other forms of research in terms of superstitious practices and beliefs:

> Learning Styles has been been thoroughly debunked. You might as well get out the Tarot cards. (Carl Hendrick, cited by Black 2018)
>
> "Open University, sort out your life. Learning Styles = Magical unicorns" (Tom Bennett, ibid)

It is our contention that these appeals to science use a flawed and stereotyped vernacular image or 'folk-view' (Thomas 2012:28) of the natural sciences for rhetorical effect but fail to probe sufficiently into the true characteristics of science. The source image captures some of the surface features of science without a theoretical understanding of scientific methods. We do not intend to be ‘anti scientific’, precisely the opposite: the intention of this paper is to demonstrate that many advocates of evidence-based teaching work with too simple a notion of 'science', with serious negative consequences. Our argument is that if appeals to science are to be used, this should be done critically, avoiding simplistic understanding of science which will constrain and distort educational research.

Limited and reductionist understandings of scientific methods

The role of experiments

In this superficial view, science is seen as consisting primarily of experiments, yet many scientific fields use few experiments; for example astronomy, meteorology, evolution - perhaps even many parts of biology. Moreover, as Thomas (2004:1-6) pointed out, experiments are generally used to verify rather than advance knowledge, and many

discoveries and inventions have not arisen from systematic procedures (eg penicillin, nylon, superconductivity, aeroplanes). Scientific method depends heavily on reflective observation, intelligent noticing, trial and error, and even intuition. We cannot, therefore, simply equate scientific methods with experiments alone. We should not neglect the diversity of evidence which science uses, nor ignore the different stages involved:

> 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).

Thomas expands on this more recently (2016:395) by pointing out that, in natural science too, we should not get fixated on 'the experimental methods of agriculture, plant science and pharmaceutical testing'. Citing Scriven, Thomas argues that 'there are many ways... to go about establishing causation beyond reasonable doubt' including 'a range of inferential manoeuvres involving trial and error, conjecture, and refutation' (p398).

> Scriven gives the examples of astrophysics, meteorology and epidemiology, where inference about cause follows observation, modelling, and calculation. He proceeds to discuss the very different domains of autopsy, geology, and engineering breakdown, where practitioners adopt similar processes of conjecture and refutation, making their way to conclusions about cause using straightforward heuristics and reasoning. (Thomas 2016:398, citing Scriven 2008:22-3)

Ontological and epistemological levels: the dangers of reductionism

It is important to make distinctions between different fields of enquiry, which partly relate to ontological 'levels'. In natural sciences, distinctions are made between physics, chemistry and biology, with physics described as 'lower' than biology. Although physical and chemical causes operate in living organisms, the disciplines of physics and chemistry are insufficient for an understanding of life processes. Similarly, in education it is possible to contrast two different approaches to a memorisation task through RCTs provided the context (students, prior learning, etc) is sufficiently stable; however the complexity of teaching in real school contexts, bearing in mind the requirement for sustainable cognitive, ethical and aesthetic development, exceeds the capacity of this form of research to determine 'what works', how and for what benefit.

Sayer (2010:5) distinguishes between the physical, chemical, biological and social levels, with a possible difference in level between psychological and social events. He condemns reductionist attempts to investigate 'higher level' phenomena through 'lower level' scientific methods. This does not mean that we can ignore the 'lower levels', as physical or chemical powers do not cease to operate in biological or social events.

> We intervene in agriculture at the physical, chemical and biological levels, through digging, watering, fertilizing, and weeding, and so on. We cannot intervene merely by thinking about agriculture. (Sayer (2010:18)

Stephen Rose's book *Lifelines* (2005) analyses reductionism in biology. He does not propose abandoning experiments, but places serious warnings:

> 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)

Steven Rose and Hilary Rose (1976:96-111) are equally concerned 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.

It is, therefore, fundamentally *unscientific* to try to explain social phenomena by applying the wrong level of scientific methodology. As we will argue, the use of RCTs in the field of education is not only difficult, it is frequently unhelpful and misleading. Further, to try to generate theory by amalgamating the results of disparate RCT-style experiments and generating an average 'effect size' is fundamentally flawed. Thomas makes this point about

education, arguing that 'each procedural domain in every science is highly peculiar, depending on its subject's form and texture' (2012:28).

Emergence and open systems

This epistemological problem has ontological roots. Although the principles of Newtonian mechanics also apply to living beings including humans (we have weight and are subject to gravity, for example), they are insufficient to explain features specific to living beings, let alone specifically to humans. Life processes entail biochemical reactions, but biochemistry is insufficient to deal with various forms of emergence such as growth and development; the reciprocal interactions between organisms and environments; and evolutionary change.

These forms of emergence also apply to human life, but there are further issues when considering causality in social studies: the fact that we are semantic, reflective and social beings with extensive (though not unlimited) powers to reshape our environments generates new forms of emergence beyond those affecting sheep and frogs. Bhaskar, in extending his theory of Critical Realism to human sciences and causal explanations, argues that social structures have powers that are not just aggregates of individual actions. His succinct illustration is that an army is not just the plural of *soldier* (Bhaskar 1979:34) but depends on structures and purposes. Although individual activity is needed to sustain and reproduce social structures, the structures inherent in our societies and cultures predate us as individuals: 'The social structure... is always already made.' (p42)

As Bhaskar (1978; 1979) argues 'closed system' explanations are rarely adequate in the natural world, and certainly not in social situations. Multiple forces are at work which may contradict as well as reinforce each other, and it is rare that a single causal factor or force may explain much of what occurs. We also need to consider human volition, habit, interpretation and interaction. Consequently, a typical mode of physics experiment or indeed RCT, based on the principle of artificially creating a closed situation by stabilising all but two variables, is almost impossible to create in social sciences. This has various consequences, including the need to reject research methods which cannot handle social complexity, and to recognise the limited predictability of educational processes and learning.

Consequences

The failure to recognise key human characteristics (agency, volition, intentionality, understanding, reflection) leads to pseudo-science which both assumes and promotes less

than human behaviour. The classic case is so-called 'scientific management' or Taylorism, essentially a technology of control to speed up factory production lines.

> The answer for Taylor was quantification: he used time-and-motion studies to develop tables of how many times the average worker could perform a given action in a given time, which could then be used to set extensive numerical targets for employees. This method depended on breaking down tasks as far as possible into simple, repetitive actions... The worker's creative process, a defining characteristic of what makes him or her human, is supressed by the need to comply with a template laid down elsewhere. (Umney 2018:66)

The Taylorist reconstruction of work as alienated labour is echoed in pedagogic situations by Pavlovian or Skinnerian behaviourism. This 'scientific' psychology assumes a faithful resemblance between the learning of captive animals and that of free human beings. It sidelines curiosity and cognition, and substitutes simplistic mechanistic stimuli of reward and punishment for richer forms of mediation through cultural tools such as language, as in Vygotskian theory. As with Taylorism, this assumes and produces a less than human form of activity - alienated learning. The model is inadequate to describe the situation.

The danger is that, through the drive to make educational research more 'scientific', pupils are represented as de-personalised data, described through a set of labels based on measurable characteristics. The dynamics of pedagogic interaction are translated into discrete 'interventions'; the complex lives of young people disappear as they become 'average' recipients of learning. The kinds of questions which can be asked or the problems which can be addressed are restricted to ones which can be answered through the approved methods and the 'scientific' model has become inadequate for the situation. The language of 'scientific research' is performative and not simply descriptive.

Furthermore, there are problems with calculations based on activities which cannot be measured directly. Attainment is not like volume, or progress like length or acceleration. They can only be input into calculations through proxy measures, which give an incomplete picture and may be problematic in terms of construct validity. Repeatability is problematic in educational settings, as the other things occurring in pupils' lives cannot be simply pushed aside.

Finally, by privileging a one-dimensional mode of research which eschews the need for theory, teachers are effectively discouraged from the kind of pedagogical reflection on 'what

works' for their *particular* pupils, drawing on theory and considering carefully factors which might be inhibiting progress.

Randomised controlled trials in social and educational settings

The use of RCTs, as social science's analogy to experiments, is the basic building block of evidence-based practice. The method has been adopted from drugs trials and other forms of RCT in Medicine with insufficient consideration of the difference between the fields. Since this has argument has been presented elsewhere, it is sufficient to summarise some core problems here:

i) Rigorous 'double blinding' provides important social protection in drugs trials. Its purpose is to remove the influence of human subjectivity and volition on the part of doctor or patient, as well as the power of pharmaceutical companies to influence results. This is impossible in education. it is impossible to alter practice without teachers and students noticing. This marks a fundamental difference between drugs trials or laboratory experiments and RCTs in educational settings. Whereas drugs trials try to eliminate the human factor because 'human volition is seen as a contaminator' social change is brought about *through* the human agent (Pawson 2006:27). This lays educational RCTs wide open to expectancy effects and Hawthorne effects (Thomas 2016:404).

ii) Sample choice generally operates in drugs trials to eliminate interference from other possible causes. This is not always productive: for example, drugs for heart disease are tested out on middle aged men without other ailment, whereas real patients tend to be older, men and women, and with comorbidity (comment from Lehman, an experienced GP, cited in Greenhalgh 2016). It is unimaginable to eliminate the 'interference' of diverse human characteristics from learners. Similarly, randomisation is rare in education RCTs, since school classes are generally pre-formed.

iii) In drugs trials the control group is typically provided with a placebo, in order to judge the relative impact of the intervention. This presents immediate problems of ambiguity for education: is the control group to do nothing in place of the intervention, or 'business as usual'. To take a simple example, in a trial concerning open questions, should the control group have entirely closed questions or should the teacher attempt to follow their normal habits. As Pawson (2006:51) puts it, 'This is not the world in repose. This is no vacuum... Control groups or control areas are in fact kept very busy.'

All of the above procedures serve to regularise and simplify in order to isolate the impact of a single 'intervention' whilst keeping everything else constant. They are the analogy to the procedures in laboratory experiments which stabilise other variables; they transform the openness and complexity of real situations into a closed system. Experimental procedures seek to de-activate other forces which might negate, distort strengthen or weaken the factor under investigation. Of course, some open systems in nature (eg the weather) are beyond the power of experiments to close and simplify. This is arguably the case with most social situations (Pawson 2006:18) including educational ones (Biesta 2010:496).

Rather than considering RCTs to be the 'gold standard' of research, we need to consider the frequency with which problems occur and are not resolved. In a study of mathematics curriculum RCTs, Ginsburg and Smith (2016:ii) identify 12 different threats to accuracy and usefulness. These include:

- authors having an association with the curriculum's developer
- curriculum interventions being poorly implemented - especially in the first year of implementation
- a failure to identity the comparison curriculum experienced by the control group
- more instructional time for the treatment than the control group
- a failure to evaluate longer term impact
- assessment tools which favour the content of the curriculum package being assessed.

They conclude that 'the magnitude of the error generated by even a single threat is frequently greater than the average effect size of an RCT treatment.' This is not an imaginary problem, since Ginsburg and Smith, in their analysis of 27 mathematics RCTs rated highly by the What Works Clearinghouse (WWC), found that 26 showed multiple problems which were sufficiently serious to make them unreliable (ibid:ii).

Empiricism, reductionism and Critical Realism

At this point it is pertinent to probe more deeply into the nature of natural sciences, in order to understand the flaws in how 'evidence-based teaching' works views science: in other words, although the appeal to science suggests rigour and objectivity, it reflects only superficial features of natural scientific methodology and principles.

The implicit assumption of many education-based RCT studies, as commissioned by EEF, is empiricism. There is scant regard for questions of causality. As an example, a major report on action to remedy literacy problems in 11-12 year olds (Gorard et al 2016) does not ask what

those problems might be; it is framed entirely in empirical measurement of the effectiveness of a particular program; the concern is simply to identify 'what works' without really grasping why. Moreover, in reporting source research projects and systematic reviews, the EEF summaries and technical appendices effectively hide from view any discussion of causality. Whilst to some advocates of the 'What works' philosophy this may seem adequate, it is hardly a scientific approach. Natural sciences are built upon the development of theory, with a reiterative interplay between observation, explanation and testing. This crucially requires an understanding of causal powers, predictability and generalisability and not merely regularity.

The tacit assumption is that regularity is as far as one needs to go in pursuit of 'what works'. Despite Hume's (1748: section VII) insistence that repeated occurrences can never establish causality, pragmatically he was content to act as if they did, and assumed that science was basically trustworthy. Similarly, despite the general agreement among statisticians and social scientists that 'correlation does not imply causality', the term 'effect size' suggests the opposite; indeed many research reports of this kind speak casually of X 'having an effect on' Y, despite the absence of a causal explanation or even the requirement for X to occur before Y. This problem is addressed strongly by Gorard and See (2013:22)

> One of the most noticeable themes from conducting a series of research syntheses is how frequently research reports use strong causal terms to describe their findings, without any apparent justification. Abbot (1998, 149) complained that 'an unthinking causalism today pervades our journals', because correlation, pattern or even opinion was too often described in strong causal terms... a major problem is authors mis-describing correlations as causal, through forgetting that statistical modelling, including multi-level modelling, structural equation modelling and path analyses merely find sophisticated correlations.

The same authors argues that four criteria are need to establish the feasibility of a causal model:

- repeated association - the association must be 'strong, clearly observable, replicable and it must be specific to X and Y'
- sequence (X must always precede Y), and 'the appearance of Y must be safely predictable from the appearance of X)'
- measurable linked changes - 'an intervention to change the strength or appearance of X... also strongly and clearly changes the strength or appearance of Y'

- a coherent mechanism - 'the simplest available without w hich the evidence cannot be explained'. (Gorard and See 2013:4, quoting Gorard 2013).

Bhaskar and researchers in the Critical Realist tradition also insist on causal mechanisms, though they are more doubtful about the above appeal to regularity, replication and transparency. Their model is based on a frequent disjunction between the 'real', the 'actual' and the 'phenomenal'. We may not experience or observe what happens, and moreover underlying forces (the 'deep real') may fail to actualise in open systems. Thus, causal forces belong to the 'deep' reality, and may be triggered or blocked by other forces or by aspects of the environment in which they attempt to materialise (see for example Bhaskar 1978, Sayer 2000:10-19).

Consequently measuring regularities is insufficient for science; attentive observation is an essential part of the process of looking beyond surface features.

> 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)

Similarly Hammersley (2015:4) makes the point 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 should note that medical research is saturated with theory:

> 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)

In the natural sciences, scientific theorising operates in conjunction with various forms of experimentation and observation to lead to reliable causal understandings.

> Natural scientists don't slavishly pursue methodological formulae about establishing

> causation. They infer, based on their prior knowledge, their critical observation, their conjecture, and their testing of these conjectures, in a process that has come to be called by philosophers of science *inference to the best explanation*. (Thomas 2016:406)

We can contrast this with an explanation of scientific methodology from Tom Bennett, a well-known advocate of RCTs and meta-analysis who appears to regard them as the only valid 'scientific' form of educational research. Bennett seems to regard experiments as a faithful reflection or re-enactment of reality, and portrays the process of designing experiments / trials as moving along smoothly from data or casual observations without theory, or indeed anything more than a superficial sequential hypothesis.

> 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)

The role of theory in scientific work is trivialised in this description. The importance of this quotation, in the present conjuncture, is that Bennett, as founder and organiser of the ResearchED conferences, has considerable influence among teachers and with government.

Statistical synthesis

Natural science is cumulative but not in a simplistic sense of piling up data. It depends on the ability to construct coherent bodies of knowledge. This requires not only gathering a body of evidence to support (or refute) key ideas, but also critical challenges to dominant paradigms

when the evidence is inconsistent or contradictory. Scientists do not operate by averaging the results of multiple experiments.

Evidence-based teaching also attempts to build a body of knowledge. It depends not only on RCTs, as the first level of research, but on the synthesis of multiple primary studies. However rather than a critical review of available research studies, it relies on a particular form of review known as meta-analysis. Although the best do include a critical review of the source studies, the core procedure, in most education studies at least, is based on averaging the effect sizes of the primary research studies. Although the starting point for this involves a process of selection, the central procedure consists of calculating a mean effect size, albeit sometimes with weighting. Effect size (ES) is assumed to be a measure of how much more effective the treatment or intervention is than the control group's experience. Effect sizes are standardised by dividing by the standard deviation.

There are many problems with these assumptions. Indeed Simpson (2017, 2018) challenges the very concept of effect size in this context, pointing out that it is an indicator of how well a trial is designed to make an effect visible, not the effectiveness of the intervention. He points to three major problems:

i) there is a lack of clarity about the control group's activity (as above)
ii) research based on a limited population (eg 11-year-old boys with reading difficulties) reduces the heterogeneity of the sample and consequently magnifies the 'effect size'
iii) using outcome measures closely related to the nature of the intervention magnifies the effect size, compared with a more general assessment tool.

Such problems only get worse when the mean effect sizes of multiple meta-analyses are compared with one another. Hattie's Visible Learning project is undoubtedly the best known example globally. His statistical calculations have been subject to powerful critique (eg Bergeron and Rivard 2017; Snook et al 2009; Brown 2013; Orange 2014; Literacy in Leafstrewn 2012; with a comprehensive guide to other critiques in Lilley 2016). To summarise some key points:

- no account is taken of the length of each intervention
- nor of the tendency for average effect sizes to reduce as children get older
- sometimes Hattie uses effect size to mean 'compared to a control group' and elsewhere to mean 'as compared with the same students before the study started'
- there are frequent doubts about the directionality of influence between factors

- studies measuring the effect on grades are mixed in with studies about the percentage of students graduating
- the impact of interventions on self-esteem, as estimated by the students, are mixed in with studies measured by attainment.

Similar problems occur in the EEF Toolkit, which uses a similar procedure but is presented within a different 'shell', namely a list of 35 types of intervention each with three pieces of data:

- mean effect size translated into 'additional months of progress'
- cost
- strength of evidence.

This enables the list to be re-sorted in various ways, though the most obvious is by months of progress.

Essentially these aggregations of research (meta-analysis and meta-meta-analyses) are less than scientific. It should be noted that many of the syntheses which the EEF draws on combine statistical averaging with a verbal review of source studies, and some of the discussion is well theorised, but unfortunately most of this disappears in the EEF reports. Some of these problems are acknowledged in the EEF Toolkit's Technical Appendices (Higgins et al 2012), though the school heads that this is meant for are likely to be too busy to notice.

Perhaps the most important problem with both the Toolkit and Visible Learning is that the selection process for source documents takes place on technical grounds, without seriously considering underlying theories, the context, or indeed whether the interventions are sufficiently similar. Radically dissimilar studies are often aggregated to produce a mean effect size - a problem known in the literature as 'Apples and Oranges'. There is little regard to differences of context (students' ages, curriculum areas, prior attainment levels etc.) This goes against the warning from Gene Glass, originator of the idea of meta-analysis, concerning 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* (Glass in Robinson 2004:29, my italics).

Sports participation: a case study

To illustrate some of these problems, let us consider, as a case study, the EEF's report on the impact of *sports participation (EEF 2018)*. (This is just one of many possible illustrations. See, for example, Wrigley 2018 for comments on *Feedback* and *Teaching Assistants*, or the many other sections referred to in Simpson, 2017.) Admittedly the notes in the Sports Participation section acknowledge the extreme variability of effect sizes in the source research, yet somehow an average is derived.

As previously explained, at the Toolkit's top level, we see a table which can be arranged in terms of *Impact* (expressed as *Additional Months of Progress).* The table ranges from +8 months for *Feedback* to -4 months for *Repeating a Year*. *Sports Participation* is given an impact score of +2 months, summarised as 'low impact for moderate cost'.

Drilling down a level, we find that 'the overall impact of sports participation on academic achievement tends to be positive but low'. It should be noted that this is an instrumental perspective, i.e. the impact which sports participation has on academic attainment (exams, test scores). The value of sports participation as enjoyment, or indeed other possible benefits such as fitness or fair play, is not considered. Sports participation might also have an indirect impact on achievement by promoting a positive ethos within a school, but this would not register in RCT-style studies with pre- and post-tests in academic skills over a relatively short time frame. Despite the low average impact of 2 months, an exception is allowed:

> Sports participation can have a larger effect on, for example, mathematics learning when combined with a structured numeracy programme (with one study showing an impact of up to ten months' additional progress). In this circumstance the 'participation' acted as an incentive to undertake additional instruction.

Further detail can be found by drilling down further to the Technical Appendix. The EEF toolkit summary effect size for sports participation (0.17, roughly two months) comes from averaging four effect sizes, derived from three meta-analyses. It would appear as if the EEF has undertaken some weighting to obtain this average (since the unweighted average is 0.31) but this is not made clear. For one of the three meta-analyses (Newman et al 2010) two different effect sizes are supplied, one for ‘academic outcomes’ and one for ‘mathematics’, with no explanation for why these two were selected from a longer list. Thus the procedures by which meta-analyses are combined to give a meta-meta-analytic mean effect size are opaque.

The figure of 0.17 derives from

0.10 (Lewis 2004)

0.15 (Shulruf 2010

0.19 for Academic Outcomes and 0.80 for Mathematics (Newman et al 2010).

Only the most determined reader, among busy headteachers, is likely to download the original research review by Newman et al to discover that the 0.80 refers to one specific sports-related intervention.

We need to drill down a further stage to the original research report to get a glimpse of why it might have been so successful. We discover that the highly successful program reported in Newman et al (2010) is *Playing for Success.* This initiative involved establishing study support centres at prestigious soccer grounds (Sharp et al 2003). Strictly speaking it was not an RCT as participants self-selected by volunteering, and then teachers decided which pupils should go forward. Underachieving pupils from local schools spent a total of 20 hours at the sports club. They enjoyed a boost to status and self-esteem through meeting star players, visiting the club's museum and boardroom, etc. The intervention was well resourced, including one-to-one mentoring and dedicated ICT suites. They had a personalised curriculum adapted to their individual needs in literacy, numeracy and ICT skills. Moreover, it was designed round practical and situated activities which were meaningful to the pupils: 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.

Playing for Success was highly successful not only for *maths* but for literacy and ICT skills, although the additional progress in maths had the edge. Although the program only lasted six months (a total of 20 hours) for each pupil, the upper primary school pupils averaged 15 months progress in reading and the early secondary pupils averaged 8 months. In numeracy the results were even more startling - over 20 months for some cohorts. By contrast, we are informed, the matched 'control group' pupils typically made *no* progress, *slipping further behind* the achievement expected for their age.

Interviews showed how the pupils had previously tended to get stuck with basic concepts at schools, lost confidence and stopped trying. When participating in the scheme, on the other hand, they became more successful independent learners, and used resources to meet their particular needs. The research evaluation reports show that they were highly motivated by the

football club context and felt 'privileged to be chosen to participate, rather than stigmatised as in need of extra help. Pupils who attend are given an opportunity that makes them the envy of their friends.' Positive attitudes were sustained a year later (Sharp et al 2003:113-120).

It is our contention that this level of causal explanation is needed if research on 'what works' is truly to inform school development. It is no use simply ascribing success to ‘sports participation’ as an incentive to receive further instruction in maths', as in the EEF summary.

A further look at the three meta-analyses drawn on by EEF (Newman; Lewis; Shulruf) shows extensive potential for the method of synthesising research to mislead. Newman et al (2010) reports the above programme (Sharp et al 2003) as one of six on sports participation, but four of the other five are problematic or irrelevant. One of them had a sample of only 15 pupils; two focused on the benefit of martial arts when compared to traditional school sports, not the benefit of sport participation over non-participation; one is about self-discovery through massage and yoga.

The meta-study by Lewis (2004) is a US-based doctoral dissertation comparing a variety of forms of extracurricular activities. One of these groups is *Sports combined with Cheerleading*, reporting a mean effect size of 0.1. The report carries a warning about 'self-selection bias':

> It is difficult and dangerous to make assumptive statements about the benefits of participation if the children who are involved in activities are so fundamentally different from those who do not. Pre-existing differences, rather than the influence of participation, may account for the gains in social competence, esteem, and achievement. (p85)

The third meta-study (Shulruf 2010) covers a wide range of extracurricular activities (ECA) and not only sports. Its warnings are even more explicit. Indeed its main aim is not to measure the impact of sports participation (and other extracurricular activities) on attainment but to *critique the assumption that a causal relationship has been found*. Many of its sources rely on secondary data; they compare retrospectively the average attainment of pupils who engaged in extracurricular activities while at school with those who didn't. The report's major conclusion reaffirms the distinction between correlation and causality, and proposes a set of criteria for determining causality, including whether multiple causes might be at work, whether the association operates under different conditions and whether there is coherence with current knowledge. Applying these to extracurricular activities, it points out that:

> ... Overall, although some associations could be identified between participation in ECA and a number of students' outcomes, there was no robust evidence for causal effects relating to these associations. Until causality can be shown, how best to enhance positive school-related outcomes through ECA will remain unclear. (p607)

The final conclusion reads as follows:

> The results show associations rather than causation and raise major concerns regarding the validity of some of the data and analyses used in the literature. This leads to the conclusion that the current knowledge on ECA participation does not suggest that extracurricular activities affect student educational outcomes either positively or negatively. It is therefore considered essential that further research be carried out to unravel how participation in ECA contributes to students' outcomes and why. Such research should investigate aspects of participation including what motivates participation, how and why students participate, and how such participation impacts on their outcomes. (p609)

This raises major questions about the 'scientific' nature of statistical meta-analysis and meta-meta-analysis. Calculating mean effect sizes is not a substitute for investigating causal mechanisms and the environments in which they activate (or don't).

The above example provides a good illustration of the conclusion drawn by Pawson:

> 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 progress 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. (2006:42-3)

Conclusion

It might appear that the discussion in the preceding section is petty and quarrelsome. Such a detailed analysis is only necessary because of frequent claims that the procedures in statistical syntheses such as Visible Learning and the Toolkit are 'scientific'. Indeed, the numerical presentation of results as effect sizes or additional months of progress creates the illusion of reliability and accuracy.

There are various dangers in the argument presented in this article. It could be read to suggest that research is of no practical benefit to teachers, because they have sufficiency of evidence gained from their own practice. This is a mistake, since research, including RCTs, can serve to challenge established professional habit. It can also highlight alternatives, though there are many forms of research which do that better than RCTs and meta-analysis (Wrigley 2018:16; Pawson 2006:50).

RCTs can be conducted without abandoning the search for theoretical and causal explanations. However, to do so they have to transcend an empiricist and instrumentalist 'what works?' mentality and engage in the classic scientific question 'What is going on?' (Rømer 2014:112). Similarly, it is important to hold onto a broader, more explanatory and exploratory tradition of evaluating a wide body of evidence through research reviews, and which select source studies on the basis of theory and relevance and not just technical conformity. (Indeed, the discussion sections of the reviews which the EEF Toolkit draws upon for its statistical data are often a better guide than its aggregate effect sizes. An excellent example is the discussion contained in the review by See and Kokotsaki (2015) cited within the Toolkit's *Arts Participation* strand.)

Essentially RCTs and meta-analyses tend to take at face value the empirical data, regarding this as sufficient for causal explanation. It is important to note Bhaskar's (1978) critique of *empiricism*, a critique which requires us to hold a distance between the phenomena we perceive and 'deep causes'. In open systems, many factors are at work which may activate or negate each other; deep forces may not actualise or become perceptible. If we are to move beyond rhetorical appeals and take scientific enquiry seriously, research literacy must extend beyond registering outcomes and actually engage with the *complex* and *situated* powers which may produce those outcomes. Teachers need to be engaged with research processes in their fullness, including focused observation, hypothesis, critical reflection on data, and clarification of aims and concepts.

At the risk of stating the obvious, educational research methods have to keep in mind the *nature and purpose of education* itself (Thomas 2012). Biesta powerfully argues that 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 phenomena.

Rømer (2014:114) relates Denmark's sudden policy conversion to 'evidence-based teaching' to a wider configuration of globalised neoliberal education governance, in which 'rankings are supposed to provide information to the global marketplace'. Thus evidence becomes a 'member of a family of concepts surrounding and aiding the processes of global capitalism.' Citing Pedersen (2011:172):

> For the first time in more than 160 years of school history, the school does not have as its primary task the formation of the individual as a citizen or a member of a democracy, but instead, the instruction of the pupil as a 'soldier' in the competition among nations. The school must now primarily promote a notion of individual competition, and is only secondarily based on the ideals of a more democratic society.

Whereas in drugs trials success criteria tend to be unidimensional and relatively unproblematic (e.g. pain reduction, a definitive cure, survival), education is marked by a multiplicity of aims - acquisition of factual knowledge, skilled performance, problem solving, longer-term cognitive development, aesthetic or ethical qualities, socialisation. An insistence on 'evidence' in the sense of numerical data (effect sizes) easily leads towards the neglect of most of these in favour of more easily measured ones such as factual knowledge acquisition and lower order understandings. A tight view of 'evidence' risks distorting curricular decisions and pedagogic practice, and abandoning such values and aims as world citizenship, multiculturalism, enlightenment, democracy, solidarity, character, virtue, knowledge and Bildung (Rømer 2014: 115).

Ironically, in the English situation at least, organisations such as ResearchEd and EEF, in their different ways, claim to empower practitioners through research literacy, yet promote and perpetuate a one-dimensional view of research, whilst marginalising broader forms of

research. Data is presented in such a way that users are seduced into prescriptive interpretations of the evidence available. These movements promote simplified analyses of complex data, leading users to look for simple generic solutions to complex situated problems.

There is a loop between a narrow view of research and evidence and a narrow understanding and practice of pedagogy. Carol Black (2018), in her very powerful essay 'Science / Fiction', relates the vogue for debunking wider research traditions to political and educational conservatism and a transmission model of education. Inadequate notions based on shallow understandings of 'science' are performative and reinforce narrow versions of curriculum and pedagogy. In the name of 'science' and in conjunction with the dynamics of high-stakes accountability systems, we are witnessing an anti-enlightenment closing down of ways of understanding and engaging with the world.

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

Bergeron, P-J and Rivard, L (2017) How to engage in pseudoscience with real data: A criticism of John Hattie's arguments in *Visible Learning* from the perspective of a statistician. McGill Journal of Education 52(1) http://mje.mcgill.ca/article/view/9475/7229 accessed 12 Nov 2018

Bhaskar, R (1978) *A realist theory of science*. Hassocks: Harvester Press

Bhaskar, R (1979) *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, pp 491-503

Black, C (2018) *Science / Fiction: 'Evidence-based' education, scientific racism, and how learning styles became a myth*. http://carolblack.org/science-fiction/ (accessed 24 August 2018)

Brown, N (2013) Book review: Visible Learning. *Academic Computing* blog, 5 August. https://academiccomputing.wordpress.com/2013/08/05/book-review-visible-learning/ accessed 12 Nov 2018

EEF (2018) Sports participation. https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/sports-participation/ (accessed 12 Nov 2018)

Ginsburg, A and Smith, M (2016) Do randomized controlled trials meet the 'gold standard'? A study of the usefulness of RCTs in the What Works Clearinghouse. American Enterprise Institute http://www.aei.org/wp-content/uploads/2016/03/Do-randomized-controlled-trials-meet-the-gold-standard.pdf (accessed 12 Nov 2018)

Gorard, S, Siddiqui, N and 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.

Gorard, S and See, B H (2013) Overcoming disadvantage in education. London: Routledge

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

Hammersley, M (2015) Against 'gold standards' in research: On the problem of assessment criteria. DeGEval conference, Saarbrücken https://www.degeval.org/fileadmin/users/Arbeitskreise/AK_Methoden/Hammersley_Saarbrucken.pdf (accessed 12 Nov 2018)

Hattie, J (2009) Visible learning: A synthesis of over 800 meta-analyses relating to achievement. London: Routledge

Higgins, S, Kokotsaki, D and Coe, R (2012) The teaching and learning toolkit: Technical Appendices. https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Technical_Appendices_(July_2012).pdf accessed 24 August 2018

Higgins, S, Katsipataki, M, Kokotsaki, D, Coleman, R, Major, L and Coe, R (2014) The Sutton Trust - Education Endowment Foundation Teaching and Learning Toolkit. London: Education Endowment

Foundation https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/ (accessed 12 Nov 2018)

Hume D (1748) *An enquiry concerning human understanding*. Republished in D Hume (1975) Enquiries concerning human understanding and concerning the principles of morals (3rd edition), ed. P Nidditch. Oxford: Clarendon Press

Lakoff, G and Johnson, M (2003) *Metaphors we live by*. (2nd ed.) Chicago: University of Chicago

Lewis, C (2004) The relation between extracurricular activities with academic and social competencies in school age children: A meta-analysis. PhD thesis, Texas A&M University. http://oaktrust.library.tamu.edu/handle/1969.1/2710 accessed 24 August 2018

Lilley, G (2016) Peer reviews of Hattie's Visible Learning (VL). www.visablelearning.blogspot.com (accessed 12 Nov 2018)

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

Literacy in Leafstrewn (blog) (2012) Can we trust educational research ("Visible Learning": problems with the evidence). 20 December. http://literacyinleafstrewn.blogspot.co.uk/2012/12/can-we-trust-educational-research_20.html accessed 12 Nov 2018

Newman, M, Bird, K, Tripney, J, Kalra, N, Kwan, I, Bangpan, M and Vigurs, C (2010) Understanding the impact of engagement in culture and sport: A systematic review of the learning impacts for young people. EPPI-Centre, Institute of Education, University of London (July 2010) http://discovery.ucl.ac.uk/1472756/1/Understanding_the_impacts_main_report.pdf accessed 24 August 2018

Orange, O (2014 a) The age effect which means the 'effect size' is useless. Ollieorange2 blog, 20 August. https://ollieorange2.wordpress.com/2014/08/20/visible-learning-6-age-and-the-effect-size/ accessed 12 Nov 2018

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

Pedersen, O (2011) Konkurrencestaten. København: Hans Reitzels Forlag

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

Rømer, T (2014) The relationship between education and evidence. In K Petersen, D Reimer and A Qvortrup (eds) Evidence and evidence-based education in Denmark: The current debate. Cursiv no 14, Department of Education, Aarhus University https://edu.au.dk/fileadmin/edu/Cursiv/CURSIV_14_www.pdf (accessed 24 August 2018)

Rose, S (2005) *Lifelines: life beyond the gene* (revised edition). London: Vintage

Rose, S and Rose, H (1976) The politics of neurobiology: biologism in the service of the state. In H Rose and S Rose (eds) *The political economy of science: ideology of/in the natural sciences*. London: Macmillan

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

Sayer, A (2010) Reductionism in social science. In R Lee (ed) *Questioning Nineteenth-Century assumptions about knowledge: II Reductionism*. New York: SUNY Press

See, B and Kokotsaki, D (2015) *Impact of arts education on the cognitive and non-cognitive outcomes of school-aged children: A review of evidence*. London: Education Endowment Foundation

Sharp C, Blackmore J, Kendall L, Greene K, Keys W, Macauley A, Schagen I, Yeshanew T (2003) Playing for Success: an evaluation of the fourth year. Nottingham: Department for Education and Skills

Shulruf, B (2010) Do extra-curricular activities in schools improve educational outcomes? A critical review and meta-analysis of the literature. International Review of Education 56(5), pp591-612

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) pp897-913

Snook, I, Clark, J, Harker, R, O'Neill, A-M and O'Neill, J (2009) Invisible learnings: A commentary on John Hattie's book Visible Learning. *New Zealand Journal of Educational Studies* 44(1):93-106

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 (2012) Changing our landscape of inquiry for a new science of education. Harvard Educational Review 82(1) pp26-51

Thomas, G (2016) After the gold rush: Questioning the "gold standard" and reappraising the status of experiment and randomized controlled trials in education. *Harvard Educational Review* 86(3) pp390-411

Umney, C (2018) *Class matters: Inequality and exploitation in 21st century Britain*. London: Pluto

Wrigley, T (2018) The power of 'evidence': reliable science or a set of blunt tools? *British Educational Research Journal* 44(3) pp359-376