FORUM
Volume 58, Number 2, 2016
www.wwwords.co.uk/FORUM
http://dx.doi.org/10.15730/forum.2016.58.2.237

Not So Simple: the problem with 'evidence-based practice' and the EEF toolkit

TERRY WRIGLEY

ABSTRACT There are increasing calls for policy and practice to be 'evidence informed'. At surface value, there may appear much to commend such an approach. However, it is important to understand that 'evidence' and 'knowledge' are being mobilised in very particular ways. The danger is that rather than promote a rich and lively debate about what counts as evidence, and how it can help educators, the reality is the development of a narrow 'what works' agenda which in turn imposes a 'one best way' approach to pedagogical practice.

Introduction

The call for 'evidence-based practice' appears so obviously correct. Who, after all, would wish to base teaching on a whim, or indeed on worn-out custom and practice? But like other terms which appear beyond question – think 'Intelligence', 'School Effectiveness', 'Leadership', 'Accountability', the 'Basics' – it is important to interrogate the meanings they have acquired within a neo-liberal policy framework. It is far too easy for such terms to succeed as ideologies precisely because they appear beyond question. To critique such terms is not to turn our back on evidence but to avoid a simplistic view. We need to know *whether*, *where* and *to what extent* it is valid. There is, of course, considerable irony when the slogan is uttered by politicians (e.g. Gibb, 2015), who are themselves so cavalier in their use of evidence (Alexander, 2014). Yet we cannot simply reject evidence because government ministers claim to favour it.

The appeal to 'evidence' must be understood within the parameters of a simplistic discourse about teaching. 'Evidence-based' in the present climate is all about efficiency – 'what works' – and efficiency in its turn is judged by simple

quantitative measurements. In evaluating learning, we almost always have to ask 'to what ends?', because education has multiple combined aims. We might be able to prove experimentally that beating children is the most efficient way of teaching times tables, but maths can also involve learning to solve problems and even to gain an understanding of social issues – maths for citizenship (Gutstein, 2012). We always need to ask whether particular invocations of evidence are enhancing teaching or thinning it down.

The insistence on 'evidence-based teaching' cannot be understood without looking at the context:

1. The accountability regime is insatiable in notching up demands for higher and higher standards, and ever more 'value added' (Ball, 2003).
2. The most powerful politicians are intent on controlling teachers because they are sure they know best, whilst rhetorically pretending to respect and indeed liberate them (Gibb, 2015).
3. The hegemonic form in which knowledge is produced and shared/imposed is in terms of numbers, which are presented as objective, unmediated, unbiased and scientific (Power, 1997; Poovey, 1998; Ozga & Lingard, 2007).
4. The fast-track teacher training routes favoured by the Government have created a need for 'teach it by numbers' recipe books (Manzone, 2016).

There is also, however, a call for 'evidence' from within the profession – and significantly from some of those who have been fast-tracked into positions of power and influence. Most loudly, Tom Bennett's popular ResearchEd conferences are built on the premise that the only important research is that which shows 'what works', and that the only reliable methodology is the randomised control trial (RCT). Bennett quite rightly condemns fads such as brain gym and VAK (Visual Auditory Kinaesthetic) as pseudo-science, but appears to think they have come from teacher educators in universities rather than commercial marketing of CPD to schools which no longer have guidance from experienced local authority advisers. He is right to criticise poor writing such as papers which confuse facts and values, or analysis which outreaches the data, but wrong to think that RCTs are the only sound methodology (Bennett, 2013).

The Strengths and Limitations of RCTs

RCTs are regarded as the building block of 'evidence-based practice'. They have been adopted as the 'gold standard' by the medical profession as a defence against the distortion of research by business interests, and particularly the pharmaceutical industry. In that field, there is a rightful insistence nowadays on randomly selected control groups, the use of placebos in drugs trials, and 'double-blind' experiments (i.e. neither the staff administering the treatment nor the patients are aware of which group is which). This has not, however, put an end to commercial distortion by Big Pharma:

> The first problem is that the evidence based 'quality mark' has been misappropriated and distorted by vested interests. In particular, the drug and medical devices industries increasingly set the research agenda ... By overpowering trials to ensure that small differences will be statistically significant, setting inclusion criteria to select those most likely to respond to treatment, manipulating the dose of both intervention and control drugs, using surrogate endpoints, and selectively publishing positive studies, industry may manage to publish its outputs as 'unbiased' studies in leading peer reviewed journals. (Greenhalgh et al, 2014, pp. 1-2)

There is no easy transfer of RCTs to school situations for several reasons:

1. Children are already in classes, and cannot usually be randomly allocated.
2. It is difficult to alter practice (for example, by asking more open questions) without either the teacher or the learner noticing.
3. There are ethical problems in the 'non-treatment' of control groups.
4. The classic experimental method involves freezing other factors than the independent and dependent variable, but children don't freeze easily.

The doctor–patient relationship is a significant factor in medical practice, but drugs or surgery will generally have an effect independent of bedside manner: the same cannot be said of teaching techniques where relationship and communication are inherent. Furthermore, whilst not denying that particular methods can have a greater or different impact, success or failure is heavily influenced by context.

What works depends on what purpose. Because of the multiple aims and consequences of teaching, it may be that a method has both positive and negative impacts in different domains. For example, forms of direct teaching could stimulate rapid memorisation but not long-term recall, flexible use of data in problem-solving, or an enduring engagement with the subject (Hattie, 2009, p. 211).

Finally, education involves more than transmission of knowledge and teachers have an ethical responsibility for shaping human beings and our social future: 'Even if we were able to identify the most effective way of achieving a particular end, we may still want to decide not to act accordingly' (Biesta, 2005, p. 1337). For example, corporal punishment might be 'effective' for learning spelling lists, but we would still avoid it.

This is not to suggest that such issues are irrelevant in medicine, where practitioners also need to be aware of conflicting effects, longer-term outcomes and considerations of patient preferences and holistic well-being and quality of life (an argument powerfully presented by Greenhalgh et al, 2014). I would argue, though, that the issues are even more entangled in the educational field.

'Scientific Method' and its Role in Educational Research

We need to move beyond the simple view of natural science and the role of experiments presented forcefully in Tom Bennett's book (2013, pp. 20-22), that experiments involve a straightforward process of isolating an independent and a dependent variable while keeping others constant. Firstly, they are not a simple extract from reality, but *artificially designed* situations constructed to make forces *visible* and *quantifiable*. This inevitably involves some distortion and a process of reduction. Steven Rose (2005, pp. 73-97) argues that scientific method often makes tactical use of simplification, but that scientists have a responsibility for reconstructing and explaining the complexity of the real world which their experiments have simplified. Secondly he points to the danger of *reductionism* in the sense of privileging 'lower-level' sciences over 'higher', pointing to fatal errors in reducing psychology or sociology to biology, or biology to physics:

> Physics deals with relatively simple, reproducible phenomena which can be measured with exquisite precision, and finds it hard to deal with complexity. Biologists' questions about the world are not easily answerable in the reduced, mathematicizing language of physics, and they are said to suffer from a sense of inferiority, of 'physics envy'… Not everything is capable of being captured in a mathematical formula. Some properties of living systems are not quantifiable. (p. 9)

Thirdly, scientific experiments do not arise out of the blue but are theory-informed. Particular procedures are designed to examine a specific hypothesis within a wider framework. Fourthly, some fields, such as the weather, human bodies and indeed classrooms, are *open systems* which operate in non-linear ways.

To elaborate an argument presented by Gary Thomas (2004), there are many scientific fields which use few experiments, for example, astronomy, meteorology, evolution – indeed biology as a whole. Experiments are used to verify, not advance knowledge, and many discoveries and inventions have not arisen from experiment or systematic procedures (e.g. penicillin, nylon, superconductivity, aeroplanes). Scientific method depends heavily on focused and reflective observation, intelligent noticing and intuition.

We should not, therefore, simply equate scientific methods with experiments.

Evidence in Teaching

This does not mean dispensing with RCTs, but recognising the limitations. It also does not mean we should disregard evidence, but we may need to reconsider what evidence means in our field.

As Thomas (2004) has argued, the nature of evidence varies in different fields: the term is used quite differently in law or history than in natural science. Here is it useful to relate to an argument pursued in medicine, namely that the practitioner's experience is as important as systematic evidence, and must be

used in conjunction with it. According to the pioneers of evidence-based medicine in Britain:

> Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research. (Sackett et al, 1996)

Trisha Greenhalgh and colleagues (2014) argue that the evidence must be individualised to the patient, and particularly where there is 'co-morbidity', such as elderly people with multiple conditions. In drawing on published research findings, the question must be, 'What is the best course of action for this patient, in these circumstances, at this point in their illness or condition?' They assert that 'real evidence based medicine is characterised by expert judgment rather than mechanical rule following' and that it 'involves finding out what matters to the patient ... Research evidence may still be key to making the right decision – but it does not determine that decision' (p. 4).

We have the added complexity in education that classrooms consist of many individuals, who may have different individual experiences and attitudes but also develop a collective response to what teachers do. What works for one class might not work for another, and what works on Wednesday morning might not work on Friday afternoon, so expert teachers have a repertoire of techniques and do not follow lesson plans slavishly.

Furthermore, aims and values are integral, not extraneous: it is not just an optional extra, or a deviation from the 'best evidence' on 'what works', to do things differently. Education for democratic citizenship is not just something we do in Citizenship or PSHE. (Here there is an interesting contrast with schools minister Nick Gibb [2015], who argued in favour of evidence-based teaching but set alongside that some of his own prejudices which he said did not require evidence: school uniform or the EBacc list of subjects, for example.)

There are implications too for education research. Tom Bennett (2013) is doubly wrong in seeking to limit this to (a) effective classroom teaching and (b) RCTs. Firstly, research must include a broader understanding of the nature of educational research:

- philosophical discussion of educational aims (what do we mean by democratic or student voice or 'British values');
- historical and comparative studies;
- the sociology of education and of children's lives more broadly;
- critical policy studies;
- ethnographies;
- the educational pursuit of social justice.

But even in terms of classroom studies, a range of methods provides insights and evidence which are difficult to construct within RCTs. For example:

- studies based on close observation/interpretation of interactions using video or transcripts;
- case studies;
- interviews with students.

These are as relevant for teacher-researchers as for academics, and it would be a mistake to promote versions of school-based research which, though rigorously designed as experiments, are lacking in philosophical or theoretical foundations and unable to perceive the complexity of classroom dynamics and teacher-pupil interactions.

Putting the 'Evidence' Together: the mystique of meta-analysis

At first sight, the term 'systematic review' appears to refer to any comprehensive and objective summary and evaluation of the available research. In current usage, it is not so innocent. The problem is that 'systematic' refers to selection criteria which are often formalistic, and operate to exclude important studies – in particular, any qualitative research and, in practice, anything not written in English.

Meta-analysis is a particular genre of systematic review which tries to calculate the average impact of a particular kind of intervention. It has all the power, and equally all the problems, indicated above. Because of the need to calculate an average, it is necessarily limited to quantitative studies, though not always RCTs; for example, it can include statistical studies. We even see arguments that badly designed studies are acceptable because errors will be averaged out.

A further step is the notion of meta-meta-analysis, the most famous of which is John Hattie's (2009) *Visible Learning* project – a meta-analysis of 800 meta-analyses, based on over 50,000 separate research studies. Apart from the sheer hubris of such a claim, we need to recognise some problematic tendencies resulting from his method and its selection process:

- the source studies are overwhelmingly from the USA, where there is enormous pressure on academics to pursue quantitative research;
- the source studies are frequently 30-50 years old;
- many of the source studies are limited by narrow outcome measures which do not reflect important educational aims, for example gap-filling, multiple choice, spelling, reading aloud single words, basic arithmetic, IQ test questions;
- a very large number are based on early literacy or numeracy.

Hattie's signature device is a dial resembling the speedometer on a car dashboard, and which expresses the mean benefit of a particular kind of intervention (Figure 1). He argues that any intervention with an effect size of less than 0.4 is not worth pursuing. Hattie calculates his 'hinge point' of 0.4 by

averaging all of the interventions in his book. Part of his logic is that the average annual improvement by students is an effect size of 0.2 to 0.4. However, this premise has been sharply questioned by other statisticians, though there is only room for a few observations here:

1. No account is taken of the duration of each intervention, which may vary from several weeks to a year or more (Brown, 2013).
2. The calculations mix together a diversity of outcomes, including literacy, numeracy, other specific curriculum areas and psychological gains, for which no norms are available (Brown, 2013).
3. No account is taken of the fact that average effect sizes reduce dramatically with age, from five to nine (Orange, 2014a, b).
4. The calculations are sometimes an aggregate of a very specific and sometimes a very broad grouping, as well as jumbling together issues such as 'home', 'personality', 'parental involvement', 'happiness' with specific teaching methods (Higgins & Simpson, 2011)
5. 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).

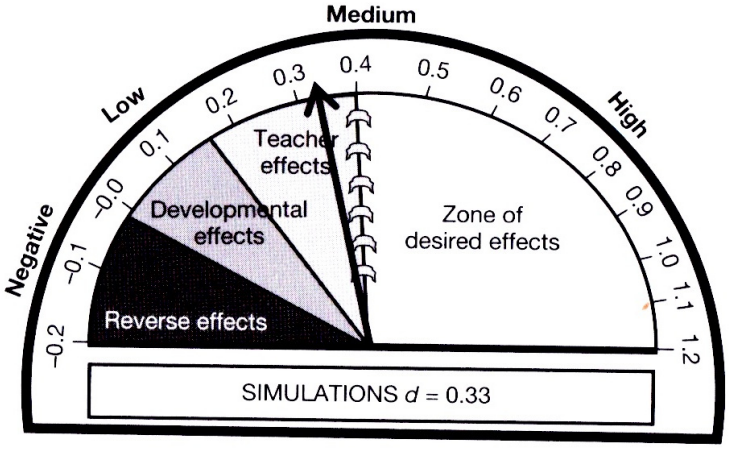

Figure 1. Example of dial from Hattie (2009).

In fact, most of the above critics conclude that Hattie is statistically incompetent, and either doesn't understand the significance of his mistakes or doesn't care. Regardless of that, his books have become international bestsellers, probably because they appear to offer hard-pressed teachers a straightforward, authoritative answer.

A central problem throughout Hattie's work and the EEF Toolkit (see next section) as 'meta-meta-analyses', along with more modest meta-analyses, is known as the *Apples and Oranges* problem. Here is Robert Coe's warning:

> 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)

Unfortunately, all these reservations were shelved by Coe and his colleagues when contracted by the Education Endowment Foundation to produce the *Toolkit*.

We can best illustrate just how misleading inappropriate averaging can be by considering the contentious issue of 'direct teaching' ('direct instruction' in US terminology). Hattie is clearly an advocate of 'direct instruction', declaring it to be 'more effective than investigative or inquiry methods', but there is considerable sleight of hand, or at least conceptual and methodological confusion, here. Firstly, he somehow manages to divorce 'direct instruction' from 'transmission teaching' or 'frontal instruction'. He concedes elsewhere (p. 209) that inquiry methods are less efficient for learning facts but *better for longer-term recall, linking concepts together, engaging students, applying knowledge, solving problems, critical thinking and scientific process!* Much depends too on whether teachers have been adequately trained in inquiry methods. Readers who are seduced by Hattie's dials, as a representation of summative data, will be seriously misled.

The EEF Toolkit

The most widespread use of a meta-analysis by England's teachers and school leadership is the *Teaching and Learning Toolkit* which was commissioned by the Education Endowment Foundation (ongoing) from CEM at the University of Durham. Following the allocation of specific additional funding for disadvantaged children known as the Pupil Premium, guidance had to be provided to schools on how best to use it and this appeared in the form of quantitative data, based on meta-analysis, or frequently (as with Hattie) meta-meta-analysis.

The Toolkit provides a rank order of around 30 different kinds of practice or intervention in terms of 'additional months of progress' (a translation of *effect size*), alongside data concerning their cost and the compiler's assessment of the strength of available evidence. No indication is given of whether this is based on a short-term remedial intervention or a long-term transformation in pedagogy or school organisation, nor of the conditions which might be necessary for it to take effect. Adding together the months of progress attributed to the 30 categories gives a total of over eight years of additional progress! These 30 categories range from specific to general, from teaching methods to school organisation to supplementary activities. The list includes, for example, phonics, collaborative learning, class size, learning styles and outdoor adventure, which presumably aren't assessed in terms of the same outcomes. All kinds of context, age group, curricular subject, and type of 'intervention' are thrown into the same pot so that the interventions can be judged on their effectiveness relative to cost. It is as if doctors were told that surgery is more beneficial than pharmaceuticals or lifestyle changes or psychiatric treatment regardless of the individual patient's needs or of whether the problem is heart disease or anxiety; or whether the surgery is a Caesarean, a lobotomy or a triple bypass. Thus, in all the ways cited above, the apples and oranges error which CEM's director, Robert Coe, highlighted earlier is now written in to the procedure and presentation and, presumably, the Department for Education contract.

There is an attempt to include further details and explanation, some of which is helpful, but the sources listed are limited, and many of them unobtainable by schools. The abstracts provided are frequently incomprehensible on a stand-alone basis, and little regard is paid to context. One exception is the division of Homework into primary and secondary school, with homework in primary schools showing up as much less effective. There is no explanation of why this might be, though by digging down further, it appears that homework is less effective in primary schools the more frequently it happens; this suggests, perhaps, that much primary school homework might be a ritual exercise to keep parents happy rather than being aligned with curricular needs.

One of the highest scorers is Feedback, though further digging reveals that the effect size is highly variable, and some kinds of 'feedback' are actually detrimental. This is hardly surprising given that the research sources on feedback range across summative and formative assessment, marks/grades and verbal explanation, comments on a completed task and oral advice during an activity; some even include as feedback a clarification of the task to the whole class. There are good reasons to believe that advice given to individuals *during* an activity can be highly beneficial, and at no cost, either financial or in teacher time. Ironically, the high scoring of Feedback led almost immediately, in the context of high-stakes accountability and Ofsted-panic, to an increase in teacher workload; *feedback* was read by many head teachers as written comments, and the practice of 'triple marking' went viral as head teachers sought to create an

audit trail on assessment for a future inspection. Thus, the Toolkit rating of 'low cost' quickly resulted in *exorbitant cost in teachers' unpaid labour*.

A further illustration of the difficulties of reducing carefully explained research into a number is provided by the low rating for classroom assistants. Again, the number merges different social contexts, age groups and pupil needs. Much of the basis for the rating was research by Peter Blatchford, who was around to speak back. His research related, in fact, to classroom assistants working in conditions where there was no time for guidance from the teacher; it was critical of practices whereby the children most in need of expert help were allocated to the classroom assistant while the more highly qualified teacher attended to others in the class. The director of the Education Endowment Foundation appears to have intervened personally, resulting in some revision of the headline rating.

In summary, whilst some information can be gained from the Toolkit which might provide a steer (including the very low rating for government-favoured practices such as performance pay, school uniform and streaming/setting), it has limited value and can be extremely misleading. Its authors are clearly aware of this, given their very wise general advice on using the Toolkit:

> 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, it seems inevitable that hard-pressed teachers and heads are likely to focus on the 'months of added progress' figures presented in league-table format, which will lead them to jump to conclusions. Most teachers will be unaware of its many problems, including that:

- the league table format encourages aggregation of dissimilar studies (apples and oranges);
- the sources are selective, and many of them contradict or qualify the headline figure;
- many interventions are context-dependent, and that much of the research supporting them is rooted in a particular usage and context;
- there is some misrepresentation of source data;
- the precise nature of interventions is invisible;
- the focus is solely on attainment, without considering other aims of education;
- much of the research does not even consider the needs of the pupils the Toolkit is supposed to help, namely those suffering from *poverty-related disadvantage*.

The Call to Follow Medicine

Great rhetorical play is made of the call for teachers to emulate doctors by espousing evidence. In the present circumstances, this is misleading, and may be intended to mislead.

Firstly, it is important to understand that the call for greater and better use of evidence emerged as a kind of social movement from within the medical professions. It was accompanied by a transformation of initial education to place a greater emphasis on the ability to read research intelligently and critically. Trisha Greenhalgh was cited earlier, and it is significant that her book, *How to Read a Paper: the basics of evidence-based medicine* (1997), now in its fifth edition, has become a standard textbook in the professional formation of doctors. One notable trend methodologically has been the adoption of problem-based learning, to inculcate an attitude of seeking sound evidence in response to specific clinical problems. This contrasts sharply with the dogmatic pursuit of fast-track routes to teaching which marginalise all theoretical learning and academic engagement. There is little doubt that it has brought benefits. Advocates argue that it has enabled young doctors to challenge habitual practices which were ill founded and even harmful. Certainly no one would suggest that doctors should *not* rely on evidence. That is not the issue.

The critique focuses on a number of interrelated issues (and here I quote from an important paper by Trisha Greenhalgh and colleagues (2014), 'Evidence Based Medicine: a movement in crisis?'):

1. It has not overcome the power of vested interests. Indeed it has been 'misappropriated and distorted' by them: they 'increasingly set the research agenda' and influence the conduct of research (e.g. 'setting inclusion criteria to select those most likely to respond to treatment' and 'selectively publishing positive studies').
2. Average results do not have sufficient fit to individual patients. In particular, 'as the population ages and the prevalence of chronic degenerative diseases increases, the patient with a single condition that maps unproblematically to a single evidence based guideline is becoming a rarity'.
3. There is no substitute for 'the subtleties of clinical judgment' and the experience which diagnostic practice builds. Doctors must take account of the complexity of each patient, and monitor the impact of treatment, as the response will differ.
4. Doctors have to consider not just the physical consequences of treatment, but what patients feel about it, and (especially but not only in the case of terminal illness) consider the needs and desires of the patient. 'Real shared decision making ... involves finding out what matters to the patient – what is at stake for them – and making judicious use of professional knowledge ... and introducing research evidence in a way that informs a dialogue about what best to do, how, and why.'

5. There are many areas of medicine which are less amenable to experimental research along the lines of drugs trials, including psychological issues and public health.
6. Doctors must resist the 'creeping managerialism and politicisation of clinical practice' in deciding on the best treatment, as well as over-prescribing which, whilst avoiding litigation, can underestimate harms.
7. Although rapid access to research evidence is key, it can also short circuit and mislead. 'Well intentioned efforts to automate use of evidence through computerised decision support systems, structured templates, and point of care prompts can crowd out the local, individualised, and patient initiated elements of the clinical consultation. For example, when a clinician is following a template-driven diabetes check-up, serious non-diabetes related symptoms that the patient mentions in passing may not be documented or acted on.'

Among the examples given to illustrate the complexity of medical practice, we find 'the 74 year old who is put on a high dose statin because the clinician applies a fragment of a guideline uncritically and who, as a result, develops muscle pains that interfere with her hobbies and ability to exercise'.

In a recent PowerPoint presentation, Trisha Greenhalgh (2016) cites a very experienced general practitioner, Richard Lehman, who has provocatively tweeted, 'Rubbish EBM = Maximally Disruptive Medicine'. Lehman has pointed out that real-life patients who come in with heart failure:

> 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 control trials cannot be applied directly to real patients.

Greenhalgh and colleagues (2014) argue not for the abandonment of evidence-based medicine, but for a return to 'real evidence based medicine'. 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.

They argue that doctors need 'a more nuanced clinical expertise that embraces accumulated practical experience, tolerance of uncertainty, and the ability to apply practical and ethical judgment in a unique case'.

Perhaps there is not such a gap as we might imagine between the fields of medicine and teaching. Certainly 'evidence-based practice' is being misrepresented in a reductionist way by those who call upon teachers to follow the medical profession, and whose real aim is not to raise the status of educators but to downgrade and de-skill them through fast-track training, draconian surveillance and teach-it-by-numbers professional guidance.

A Provisional Conclusion

This article is not a call to turn our backs on evidence, but rather to avoid a simplistic view. In particular, we must avoid the assumptions:

- that evaluating teaching involves only measuring 'what works';
- that research can show 'what works' in general;
- that experiments are the only, or necessarily the best, way to find out;
- that evidence can be compiled through meta-analysis into a league table of 'effect sizes'.

We cannot allow 'evidence' to replace a teacher's professional judgement. A teacher's experience also provides worthwhile evidence. Teachers, like doctors, rely on empathetic listening, relationships, a sense of individual difference, careful monitoring, and an understanding of complexity.

As we saw in the last section, medicine has its complexities and we should be suspicious of the rhetoric of those who cite it in their espousal of 'evidence-based teaching', but education has additional ones:

- subjective and intersubjective factors are even more crucial;
- student expectations, choices and reactions are critical to the success or failure of any teaching method;
- educational aims are unsettled and multi-layered;
- learning is non-linear.

A teaching method is only 'effective' in terms of specific aims, and what might impact positively in terms of one aim could be harmful in terms of others.

Pedagogical decisions can only be reached on the basis of pedagogical consideration and not through technical procedures. Here I am using 'pedagogical' not simply to mean 'teaching methods' but as involving key questions about the kind of human being and the kind of society we seek to develop. 'What works' should always be subservient to ethical, social and political questions: evidence should never replace judgement about the purposes of educational activity and the nature of education.

Statistics deals in probabilities, not certainties. Particular teaching methods may work for some students and not others, and indeed for some teachers but not others. A teaching method is not 'effective' in general terms, regardless of (i)

the learners' age, prior attainment, learning experiences, lifeworld; (ii) the structure and ethos of the school; (iii) the curriculum area, and indeed the nature of (iv) the outside world. There are many valuable forms of classroom research and evidence, including close observation, recording what students say and do, talking with students about their learning, and case studies which look at the complexity of a particular situation. Moreover, research should not be limited to what happens within classrooms, but involve a wider sociology of our students' lives, class structure, racism and so on; philosophical discussion of the purposes of education; curriculum studies; learning theory, and so on.

Finally, we need to ask paradigm questions about the call for 'evidence-based teaching' and the research which is said to underpin it. Simple linear cause–effect does not belong in open systems, and as Biesta (2014) argues, education is an 'open, semiotic and recursive system'. Teaching and learning cannot therefore be understood in terms of efficient *interventions* which cause *outcomes*, but as communication and interaction. In other words, teaching needs to be explored not as 'physical push and pull' but as a 'process of meaning/interpretation'.

For all these reasons, professional judgement cannot be reduced to technical calculations based, directly or indirectly, on comparative attainment scores. This is not to suggest that attainment is unimportant, but that it cannot be understood in isolation from a broader spectrum of educational aims and outcomes. I would not wish to suggest that randomised control trials and meta-analyses are never useful in education, but their limitations must be understood.

A richer, more contextualised and dialogic relationship between researchers and teachers is needed through which research can be mobilised, mediated and made accessible in the all-round interests of young people and our social future, and involving academic support for practitioner research (Beckett et al, 2014). A quick-fix, scores-based league table is no substitute, however much it is in tune with high-stakes accountability and neo-liberal politics.

References

Alexander, R. (2014) From Phonics Check to Evidence Check. Cambridge Primary Review Trust. http://cprtrust.org.uk/cprt-blog/from-phonics-check-to-evidence-check/ (accessed 17 December).

Ball, S. (2003) The Teacher's Soul and the Terrors of Performativity, *Journal of Education Policy*, 18(2), 215-228. http://dx.doi.org/10.1080/0268093022000043065

Beckett, L. et al (2014) Raising Teachers' Voice on Achievement in Urban Schools in England. Special issue of *Urban Review*, 46(5), 783-923. http://dx.doi.org/10.1007/s11256-014-0301-x

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.

Biesta, G. (2005) Knowledge Production and Democracy in Educational Research: the case of evidence-based education, *South African Journal of Higher Education*, 19, 1334-1339.

Biesta, G. (2014) Who Knows? On the Ongoing Need to Ask Critical Questions about the Turn towards Evidence in Education and Related Fields, in K. Petersen, D. Reimer & A. Qvortrup (Eds) Evidence and Evidence-based Education in Denmark: the current debate, *Cursiv,* 14. http://edu.au.dk/en/research/publications/cursivskriftserie/cursiv-14/

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

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, 12-14 September. http://www.leeds.ac.uk/educol/documents/00002182.htm

Education Endowment Foundation (ongoing) Teaching and Learning Toolkit: an accessible summary of educational research on teaching 5-16 year olds. https://educationendowmentfoundation.org.uk/evidence/teaching-learning-toolkit

Gibb, N. (2015) *The Importance of the Teaching Profession.* Keynote speech at ResearchEd conference, 5 September. https://www.gov.uk/government/speeches/nick-gibb-the-importance-of-the-teaching-profession

Greenhalgh, T. (1997) *How to Read a Paper: the basics of evidence based medicine.* London: BMJ.

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* 2014: 348:g3725 (13 June).

Gutstein, E. (2012) Using Critical Mathematics to Understand the Conditions of Our Lives: United States, in T. Wrigley, P. Thomson & B. Lingard (Eds) *Changing Schools: alternative ways to make a world of difference.* London: Routledge.

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

Higgins, S. & Simpson, A. (2011) Visible Learning: a synthesis of over 800 meta analyses relating to achievement, by John Hattie, *British Journal of Educational Studies,* 59(2), 197-201. http://dx.doi.org/10.1080/00071005.2011.584660

Higgins, S., Kokotsaki, D. & Coe, R. (2012) The Teaching and Learning Toolkit. https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Teaching_and_Learning_Toolkit_(July_12).pdf

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

Manzone, J. (2016) Factory-farmed Teachers Will Fail Our Children, *Schools Week*, 13 February. http://schoolsweek.co.uk/factory-farmed-teachers-will-fail-our-children/

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.

Orange, O. (2014a) 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/

Orange, O. (2014b) John Hattie Admits That Half of the Statistics in Visible Learning Are Wrong (part 2). Ollieorange2 blog, 24 September. https://ollieorange2.wordpress.com/2014/09/24/half-of-the-statistics-in-visible-learning-are-wrong-part-2/

Poovey, M. (1998) *A History of the Modern Fact: problems of knowledge in the sciences of wealth and society.* Chicago: University of Chicago Press. http://dx.doi.org/10.7208/chicago/9780226675183.001.0001

Power, M. (1997) *The Audit Society: rituals of verification.* Oxford: Oxford University Press.

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

Sackett, D., Rosenberg, W., Gray, J., Haynes, B. & Richardson, S. (1996) Evidence Based Medicine: what it is and what it isn't, *BMJ* 1996: 312:71-72 (13 January). http://www.dcscience.net/sackett-BMJ-1996.pdf

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

TERRY WRIGLEY is Visiting Professor at the University of Northumbria. *Correspondence*: terrywrigley@gmail.com