Australasian Journal of Special Education
Vol. 32, No. 1, April 2008, pp. 5–21

Routledge
Taylor & Francis Group

EDITORIAL ARTICLE

Why Can't a Teacher Be More Like a Scientist? Science, Pseudoscience and the Art of Teaching

Mark Carter* and Kevin Wheldall
Macquarie University Special Education Centre, Macquarie University, Sydney, Australia

In this article, the authors argue the case for scientific evidenced-based practice in education. They consider what differentiates science from pseudoscience and what sources of information teachers typically regard as reliable. The What Works Clearinghouse is discussed with reference to certain limitations of its current operation. Given the relative paucity of 'gold standard' research in education, an alternative model for assessing the efficacy of educational programs is proposed as a temporary solution.

It seems that, these days, the word 'research' has been extended to mean almost any perusal of available source material, no matter how casual the approach or dubious the source, for whatever purpose. 'Surfing the net' is commonly termed 'research', for example. But real research, of course, amounts to rather more than the passive consumption of ill-digested snippets of information from sources of unknown veracity. Real research implies a critical, direct examination of original source material, rigorous data collection completed objectively, and conceptual synthesis of what is found with what was previously known. Real research is a worthy activity if, as a result, we become more knowledgeable about the reality of our world and our place within it.

Perhaps this problem is even more pronounced when we consider what counts as real *scientific* research. Science, and its derivative adjective 'scientific', are regarded as conveying at least a patina of reliable respectability when invoked to consider the truth value of statements, propositions, controversies and claims for cures, for example. As any linguistic philosopher would have it, however, it all depends what you mean by the term 'scientific', since it is frequently hijacked for use by those whose activities may more accurately be described as 'pseudoscience'. The features

*Corresponding author. Macquarie University Special Education Centre, Macquarie University, NSW 2109, Australia. Email: mark.carter@mq.edu.au

ISSN 1030-0112 (print)/ISSN 1833-6914 (online)/08/010005-17

DOI: 10.1080/10300110701845920

that distinguish scientifically-based practices from those based on pseudoscience have been explored in detail in both popular (Sagan, 1997; Shermer, 1997) and educational literature (Sasso, 2001; Stephenson, 2004). While by no means a comprehensive consideration of this issue, some key features of a scientific approach and points of divergence with pseudoscience will now be discussed.

Science and Pseudoscience

Conservative in Interpretation

Those working from a scientific perspective are typically conservative and circumspect in their interpretations. In fact, in the face of a positive finding a scientifically orientated researcher's first response is often to point out the limitations in methodology, issues in interpretation and the need for replication of findings. Definitive, or even strong conclusions, when they bear on important practical matters, would, in general, not be drawn on the basis of a single study, even a very well-designed and executed one, unless the evidence was absolutely compelling. Unfortunately, evidence of this compelling nature is not common in education. This inherently conservative approach to interpretation often proves less than attractive to the popular press. In contrast, those with a pseudoscientific orientation have little restraint in overselling findings that are unreliable or equivocal, without reservation or caveat. More typically, interpretation of data, or more often anecdotes and testimonials, is frequently extravagant. See, for example, Stephenson and Wheldall (2008) in this issue in their critical consideration of the Dore program.

Propositions are Testable and Falsifiable

A critical characteristic of any proposition with a claim to be scientific in nature is that it is testable and potentially falsifiable: that is, a proposition can be subject to objective evaluation and can be demonstrated to be incorrect. The claim that a given intervention is effective can certainly be tested using experimental research and this would be the approach of an educator with a scientific orientation. In contrast, when such research is absent or contrary to an established belief, those subscribing to a pseudoscientific method invariably resort to anecdotes and testimonials to support their position. Beliefs based on such evidence are typically difficult to counter for a number of reasons. The evidence, typically testimonials and anecdotes, usually deals with historical information that often cannot be verified and tested in a scientific sense, and thus it cannot be falsified. It is also difficult to argue the validity of a proposition with somebody who claims to have 'witnessed something with their own eyes' because of their embedded failure to recognise the fundamentally flawed nature of human perception and memory.

The history of science is littered with examples that clearly demonstrate misperception and self-deception. One classic example occurred around the turn

of the twentieth century when a leading physicist, René Blondlot, announced the discovery of a new form of radiation, the N-ray (Huizenga, 1993). This discovery, confirmed by dozens of other scientists (Carroll, 2005), led to numerous papers as well as a prestigious scientific award for the discoverer (Huizenga, 1993). Unfortunately, somewhat mischievous sabotage of one of Blondlot's experiments led to the even more amazing discovery that he could see the effects of N-rays, even when the apparatus was disabled (Huizenga, 1993)! N-rays did not exist and the episode demonstrated that even the most disciplined of scientific minds were not immune to misperception and self-deception.

Before we laugh too heartily at the sadly misguided physicists, we should recall the facilitated communication debate of the early 1990s. Many academics, teachers and parents sincerely believed their eyes when individuals with severe disabilities produced often remarkable communication by typing messages with the aid of a facilitator guiding their hand. Very regrettably, these hopes proved to be false as controlled trials clearly demonstrated that the observed communication was usually being generated by the facilitator and not the individual with a disability (Jacobson, Foxx, & Mulick, 2005; Shane, 1994). It should be stressed that there was no evidence that the facilitators were acting fraudulently. They were as much victims of self-deception as those advocating facilitated communication. The lesson to be learned is that those who believe what they see will often see what they believe.

Controlled Research

Carnine (2000) argued that a mature profession:

> is characterized by a shift from judgments of individual experts to judgments constrained by quantified data that can be inspected by a broad audience, less emphasis on personal trust and more on objectivity, and a greater role for standardized measures and procedures informed by scientific investigations that use control groups. (p. 9)

Inherent in such maturation is a shift from reliance on beliefs, anecdotes, testimonials and in particular, expert opinions. Controlled research provides us with an objective process to test theories and methods to determine if interventions are of benefit to the individuals we are responsible for serving. It is most certainly true that experimental research involving control groups is not always possible in all areas of science or all areas of education. It should also be stressed that not all controlled research necessarily involves random assignment and control groups. When dealing with very low incidence conditions, characterised by heterogeneity across individuals and idiosyncratic responses to intervention, quasi-experimental small n studies provide a viable option in many instances. Nevertheless, research involving randomised controlled trials should certainly have a central and critical role in education. In areas relating to crucial educational outcomes, controlled trials and random assignment remain exceedingly rare (Seethaler & Fuchs, 2005).

Evaluates all Relevant Evidence

In education, we more often deal with murky applied issues where data are inconsistent and messy and where research is not of a uniformly high quality. Thus, in order to form sensible and responsible conclusions, we need a balanced consideration of all of the available evidence. Evidence is typically evaluated and synthesised with consideration of both its quality and strength. Conflicting evidence is considered, explained and reconciled where possible. The strength of our conclusions is tempered where apparently contradictory evidence cannot be rationally explained. We typically form tentative conclusions that should be revised as further evidence becomes available. In contrast, a pseudoscientific position is not constrained by the need to synthesise the entire corpus of evidence. In fact, more typically, only supporting evidence needs to be considered with little regard to its quality. The view is often implicitly taken that the plural of anecdote is research.

Mainly Evolutionary

While revolutionary discoveries in science are often highlighted, the process is, for the most part, an evolutionary one. Revolutions, like the unravelling of the underlying structure of DNA by Watson and Crick or the discovery of the role of Helicobacter pylori in the development of gastro-intestinal ulcers by Warren and Marshall, are stunning, but all too rare. More typically, small pieces of information are accumulated, tested and synthesised into our existing knowledge, leading to a gradual process of revision and adjustment. While revolutions do occasionally happen in science, they are claimed with monotonous regularity in the realm of pseudoscience. Again, see Stephenson and Wheldall (2008) for a consideration of the claims made for the Dore program.

Essentially Subversive

Science is sometimes viewed as a conservative instrument that maintains the social status quo. This view can be reinforced by the conventional scientific burden of proof, which requires those proposing a change to current knowledge to prove their case before it is accepted. This position is logical since human history clearly demonstrates that we usually explore many incorrect propositions, which fail to advance our understanding, before we identify one that is correct. Such conservatism is important since we need to be reasonably sure that when we revise our accepted knowledge, it does represent a clear advance on what came before. How, then, can a scientific approach be considered subversive?

Ultimately, if a scientific approach is taken, there is no authority except evidence and all ideas are open to challenge and revision based on evidence. Even the most authoritative and lauded experts are considered to be wrong if the weight of evidence falls against them. For example, a number of Nobel laureates offered support for cold fusion during the well-documented scientific controversy in the late 1980s and

early 1990s (Huizenga, 1993). It is deeply regrettable that, given the current concern about climate change, the cheap, inexhaustible and non-polluting energy source foreshadowed in the original University of Utah cold fusion press release (see Huizenga, 1993, pp. 289–291) has not eventuated. Similarly, Linus Pauling, a Nobel laureate in chemistry and Nobel Peace Prize winner, vigorously promoted large doses of vitamin C as an important treatment for cancer, a proposition that has not been substantively supported (American Cancer Society, n.d.; Barrett & Jarvis, 1993).

If a scientific approach to education is adopted, there are no ultimate authorities and expert opinion is just that, opinion. Authority and expert opinion, regardless of how fervently it is held, dissolve in the presence of contrary evidence, making science a fundamentally subversive endeavour. Heresy and dissonance are implicit features of a scientific approach, with evidence being the ultimate arbiter. In contrast, pseudoscience often reifies expert opinion or authority, tradition, or experience, at least when it is supportive. Interestingly, heresy and dissonance with conventional science are also often highlighted by those adopting a pseudoscientific paradigm. The distinguishing feature in this instance, however, is that these views are uninformed and 'uninformable' by evidence and, thus, cannot be falsified.

How Do Teachers Know What to Do in the Classroom?

Basic knowledge of education and skills in teaching should be established in initial teacher preparation programs. It is, however, also important that teachers continue to develop and update their knowledge. The recent moves to implement professional development requirements for teachers (e.g., NSW Institute of Teachers, n.d.; Victorian Institute for Teaching, n.d.) provide an example of the recognition of the potential importance of professional development in improving teachers' knowledge. The obvious question is: how do teachers gain this knowledge? Based on available evidence, it is probably not through reading research in professional journals. In a recent review of teachers' professional reading habits Rudland and Kemp (2004) reported that teachers engaged in little professional reading, particularly when compared to other professional groups. In addition, much of this reading involved practically orientated periodicals as compared to research-based professional journals. Landrum, Cook, Tankersley, and Fitzgerald (2002) reported that both regular and special education teachers rate the opinions of colleagues, workshops and in-service programs as not only more accessible and usable, but more trustworthy than professional journals. Further, Boardman, Arguelles, Vaughn, Hughes, and Klinger (2005) reported that when making decisions about classroom implementation of practices, special education teachers did not consider it important that they be research-based.

If Carnine (2000) is correct and a mature profession is characterised by a shift from reliance on opinion and subjective judgment to quantified data that can be inspected by a broad audience, to objectivity, and to controlled research, then the

teaching profession remains firmly anchored in the pre-scientific era. Nevertheless, there are a few signs of development. Notwithstanding the issues of whether professional development activities address research-based practices, the requirement that teachers engage in professional development in any form can be seen as a step forward. In addition, the recent attempts to develop guidelines for evaluating practice from an evidence-based perspective (e.g., Gersten et al., 2005; Horner et al., 2005; Odom et al., 2004; What Works Clearinghouse, 2006a) must also be considered a move in the right direction, although as will be discussed, much of the devil is in the detail and a heavy price will be paid if educational researchers get the detail wrong.

Does 'What Works' Work?

The move towards evidence-based practice in education has been accompanied by an increasing demand for evidence of efficacy of educational programs and interventions. Unfortunately, given the decline of scientific research in education over recent decades, in favour of more ideologically-driven approaches, such empirical evidence of efficacy is thin on the ground. For example, in exasperation with the tardiness of the US government backed What Works Clearinghouse (WWC) (2006a) in issuing recommendations of effective educational programs, the website has been dubbed the 'Nothing Works Clearinghouse' (Viadero, 2006) by its critics! Unfortunately, the What Works Clearinghouse may be rapidly moving from the 'Nothing Works Clearinghouse' to the 'Almost Anything with One Controlled Study Works Clearinghouse'. In an ideal world, only gold standard research would be considered in reviewing relevant studies. In education, we currently do not live in an ideal world. Gold standard control group studies are rare in education. The dilemma faced by the WWC is whether review should be limited to only the best studies, often resulting in only a small number, or even a single study being considered, or to look at a much larger body of evidence that may lack the highest degree of rigor. The WWC has taken the former approach but has then drawn conclusions based on very limited bodies of evidence. As previously noted, in science a conclusion would rarely be based on a single or even small group of studies, particularly when the number of participants is relatively small.

This problem is well illustrated in the case of Arthur, an animated aardvark that forms the central character in a children's television show. Based on a single randomised trial with 102 bilingual children, it received the second highest rating of 'potentially positive effects' on English language development (WWC, 2006b). Interventions are rated on a six-point scale and the only higher rating is 'positive effects' (WWC, n.d.a.). 'Potentially positive effects' refers to 'evidence of a positive effect with no overriding contrary evidence' (WWC, n.d.a, p. 1).

It is questionable whether any conclusion, even 'potentially' positive, should be made on a single study, but examination of the detail raises even greater concern. The mean effect size for the intervention was only 0.29 (WWC, 2006c). To put this

in context, effect sizes of around a quarter to a third of a standard deviation are typically considered to approach the threshold for clinical significance in special education research (e.g., Forness, 2001) and the WWC have opted for 0.25 (WWC, n.d.a). In any case, the magnitude or the effect size associated with watching Arthur is marginal at best. Further, examination of the relevant technical document reveals that the effect size for one of the three reported dependent measures was negative (treatment group scored lower) and none was statistically significant (WWC, 2006c). That is, all the differences were so small that they could be due to chance. While the mechanical processes of review may well have been followed (WWC, 2006a) and the conclusions may be consistent with the relevant guidelines (WWC, n.d.a), there is clearly a major problem with interpretation.

We would stress that we are not criticising the Arthur educational television program, which incidentally has provided many hours of enjoyment to our children. Further, it is quite possible that the program may well have very desirable effects on English language learning. Rather, the question is whether any sensible conclusion can be drawn on such limited evidence and apparently flawed interpretation. Unfortunately, this example is not isolated and the WWC website is replete with similar interpretative problems. In many cases there are just too few data to draw any sensible conclusion. Admittedly, and to their credit, the WWC has recently added an extent of evidence index (WWC, n.d.b.), which, in the most recent reports, at least signals to the educator those conclusions that are based on very limited research.

Not only is interpretation flawed, but errors are made in classifying studies. The WWC report on Reading Recovery is a good example. Out of 78 studies considered only four met the critera for inclusion (plus one 'with reservations'). An evaluation by Center, Wheldall, Freeman, Outhred, and McNaught (1995) was, in company with many other studies, said not to have met 'WWC evidence screens' for the following (specific) reason:

> Incomparable groups: this study was a quasi-experimental design that used achievement pre-tests but it did not establish that the comparison group was comparable to the treatment group prior to the start of the intervention. (p. 7)

While we cannot comment with any certainty about the similarly rejected studies, we can speak with some confidence about the design characteristics of a study which one of us (Wheldall) personally designed. This might appear somewhat defensive on the part of a disappointed researcher/author, but our point here (apart from setting the record straight) is, simply and importantly, to use this case as an example by which to illustrate that the WWC screening procedures are by no means foolproof and appear to be in need of far greater quality control. Moreover, this study also highlights another major flaw in the WWC approach, as we shall see.

It is first necessary to describe briefly the Center et al. study, described by the American reading researchers, Shanahan and Barr (1995), as one of the 'more sophisticated studies'. Here is the published abstract in full:

> The authors evaluated the effectiveness of Reading Recovery (RR) in 10 primary schools in New South Wales. Children *were randomly assigned to either RR or a control condition* in which they received only the resource support typically provided to at-risk readers. Low-progress readers from five matched schools where RR was not in operation were used as *a comparison group*. Results indicated that at short-term evaluation (15 weeks), the RR group were superior to control students on all tests measuring reading achievement but not on two out of three tests which measured metalinguistic skills. At medium-term evaluation (30 weeks) there were no longer any differences between the RR and control children on seven out of eight measures. Single-case analysis suggested that, 12 months after discontinuation, about 35% of RR students had benefited directly from the program, and about 35% had not been 'recovered.' The remaining 30% would probably have improved without such an intensive intervention, since a similar percentage of control and comparison students had reached average reading levels by this stage. (p. 241, emphases added)

As may readily be appreciated from even a cursory glance at the abstract, this study did *not* employ 'a quasi-experimental design', it used a fully randomised design where subjects were randomly allocated to experimental and control conditions. It employed a comparison group *as well as* a control group, an important factor that the WWC staff clearly missed, unlike Shanahan and Barr. Moreover, as for comparability of the experimental and control (not comparison) groups, the article clearly showed and stated that 'there were no significant differences between the two groups on any literacy measures at the pretest stage' (p. 251). It was subsequently shown that there were no significant differences between the control and comparison groups either, at any of the testing points, including pre-test (p. 252).

The other main point to emphasise from this study, which advocates of Reading Recovery who frequently cite the study often miss, is that while Center et al. clearly demonstrated the efficacy of Reading Recovery, they also report important evidence that impinges on its cost effectiveness. Center et al. showed that about one-third of Reading Recovery students would have recovered spontaneously (i.e., without intervention) while a further third were not recovered. Moreover, those who were recovered were those who were shown to be less phonologically impaired from the outset.

It should also be noted that one of the studies that *was* included in the WWC evaluation of Reading Recovery (Iverson & Tunmer, 1993) was, in fact, quite critical and successfully demonstrated how Reading Recovery's effectiveness could be significantly improved. A modified Reading Recovery group received a standard Reading Recovery program that included explicit instruction in letter–sound patterns instead of letter identification procedures. The modified Reading Recovery students learned to read much more quickly than the regular Reading Recovery students.

Thus, it can be seen that the WWC approach is clearly subject to error, in this instance at least. We should, then, resist attempts to reify WWC. They appear to be reports prepared by a team of (we are sure) very competent but essentially generalist research methodologists who cannot possibly be expected to be expert in all of the

projects/programs they attempt to evaluate. We believe that you actually do need to know something about the area you are attempting to evaluate no matter how competent methodologically you may be. Had the report on Reading Recovery been conducted by expert reading scientists, we suggest that they would readily concede that Reading Recovery works but would also conclude that it does not work very well (not surprisingly since the phonics engine is seriously underpowered), works only for a minority of students who receive it (one in three, we estimate), works only for the less phonologically challenged, and is manifestly not cost effective.

While the idea of WWC is certainly a good one, attempting to draw conclusions in the absence of sufficient studies that meet quality standards, borders on being plain silly. As previously argued, in an ideal world, we would limit ourselves to perhaps a few dozen gold standard randomised controlled trials when evaluating educational interventions. Unfortunately, very few (if any) educational interventions would even approach this standard of evidence. Rather than simply discarding the vast majority of our evidence and drawing conclusions based on minimal numbers of studies, one approach would be to examine all the best evidence that is available and weight it in terms of its quality. That is, we give better-quality evidence a higher weighting in making decisions. As higher-quality studies become available, we discard the lowest-quality evidence and revise our recommendations as appropriate. Ideally, we eventually reach a point where we are only considering gold standard studies. An alternative approach is to continue to examine all relevant evidence and to determine whether apparent intervention effects vary across study quality. Effects that dissipate with increasing study quality would be of obvious concern.

We look forward to the time that gold standard evidence in education is thick on the ground and this is the only evidence we need to examine. Until this time arrives, we still need to make decisions and may be well advised to look more broadly at the evidence that is available. Failure to do so will play into the hands of those who eschew recommendations based on scientific evidence in favour of policies that are more ideologically driven. Noting the aphorism that 'nature abhors a vacuum', any perceived gap in the market will quickly be filled by policies and practices for which there is no evidence. Given the relative paucity of acceptable scientifically conducted efficacy studies in education, perhaps there is a need to consider a more measured approach, at least initially, in our determination of the acceptability of programs and interventions.

An Alternative Model for Evaluating Efficacy of Educational Programs

When considering the many and various educational programs promoted, the evidence in their support varies considerably. Supporting evidence can be considered at a number of levels. At a most basic level, programs may be consistent with existing evidence in terms of current theory and suggested practices. It is important to note that in this context, the term 'theory' is used in a scientific sense to refer to explanations for phenomena that have been verified by repeated empirical testing

and are broadly accepted by the scientific community. Programs of this nature may be viewed as 'based on' scientific research in the sense that they are conceptually consistent with evidence but this does not guarantee that any particular implementation will be effective (Slavin, 2003). To reach the highest standard of evidence, individual programs must also be specifically and rigorously evaluated (Slavin, 2003). There are (fewer) programs that, in addition to being 'based on' scientific research, have empirical evidence for their specific efficacy; a minority of the latter can also point to true randomised controlled trials demonstrating efficacy—the so-called 'gold standard'. On the other hand, we have programs that make no conceptual sense in terms of our current theoretical understanding, advocate practices that are not consistent with scientific research evidence, have no or very dubious specific evidence for program efficacy, and even programs that are predicated on assumptions counter to substantial scientific evidence to the contrary. We might, in fact, posit a sliding scale of levels of acceptable programs, similar perhaps to the Australian Travel Advisories. The following is intended as a general guide rather than an operational model for evaluating research.

Level 1: Use with Confidence

Level 1 programs are consistent with existing scientific evidence in terms of current theory and recommended practice. In addition to being 'based on' scientific research, they are supported by a number of independent randomised controlled trials providing strong evidence for specific efficacy. This is the 'gold standard' to which all programs aspire and may be recommended with confidence. Unfortunately, they are very few in number.

Level 2: Promising

Level 2 programs are also consistent with existing scientific evidence in terms of current theory and recommended practices. Empirical evidence for specific program efficacy is more limited, however, and may not include many, or indeed any, independent randomised control trials. Thus, these programs do make conceptual sense in terms of our current research knowledge but specific supporting evidence is more limited. Evidence for such programs would typically be based on strong quasi-experimental studies, including non-equivalent control group designs with pre-test matching. This level of evidence would count as 'very promising' and such programs could be recommended with a reasonable degree of confidence. It constitutes a 'silver standard', pending the collection of stronger evidence.

Level 3: Worth a Try

Level 3 programs make conceptual sense. There is typically empirical support for the type of component interventions employed and the programs are consistent with

current theory. There is, however, little or no empirical evidence for the specific program. For example, a reading program that addressed the five elements of effective reading instruction (phonemic awareness, phonics, fluency, vocabulary, and comprehension) could be considered 'based on' research but it may also be poorly implemented and ineffective (Slavin, 2003). Clearly, there is a need for supportive empirical evidence of specific efficacy of a given program before it can be wholeheartedly recommended for wide application. Such programs may be 'worth a try' in the absence of better evidence since they at least make conceptual sense but they should be used with caution. This is the minimum basis for program recommendation and constitutes the 'bronze standard'.

Level 4: Not Recommended

These programs provide no credible specific empirical evidence, do not make conceptual sense in terms of current theoretical knowledge and typically employ component interventions that do not have empirical support. They are neither 'based on' scientific research nor supported by specific evaluation. While proponents may claim (limited) empirical evidence to support specific program efficacy, this does not stand up to even the most basic scientific scrutiny. For example, many perceptual motor programs that claim to affect reading outcomes typically fail to present credible evidence of specific program efficacy, advocate types of interventions that have been previously shown to be ineffective and are not consistent with our theoretical knowledge of reading acquisition (Stephenson, Carter, & Wheldall, 2007). Such programs should not be adopted without further substantial empirical evidence for their efficacy and do not meet even the lowest standard of acceptability. Proponents of such programs should be invited to provide specific evidence, or at the very least cite supporting generic scientific research evidence, or desist from making their claims. This is the brass standard; when highly polished it might superficially resemble gold but is soon shown not to be so, on closer examination.

Level 5: Educationally Unsafe

Level 5 programs have no credible specific empirical evidence and are predicated on assumptions counter to substantial scientific evidence and theory. While programs at Level 4 may be considered to be unproven and inconsistent with much existing scientific knowledge, those at Level 5 are closer to being disproved and are antithetical to empirical evidence. These programs are the educational equivalent of homeopathy. For them to work, large bodies of established and well-validated knowledge would need to be overturned. Such programs should not only *not* be adopted but the public should be warned that they are highly unlikely to be effective and, rather than meeting any standard, should be regarded as requiring the educational equivalent of a 'health warning'. At best this is the tin standard.

Some Examples

In order to put some flesh on the bones of the skeletal model proposed above, we might consider where currently known educational programs and interventions might be located among the levels. Reading Recovery, for example, and as discussed above, must be one of the most influential, widely known and promoted educational interventions ever but where would it sit in our levels? It makes only limited conceptual sense since it appears to have remained largely unchanged over the past thirty years or so in spite of the considerable body of scientific evidence accumulating over that period as to how reading works and is best taught. Moreover, while there is a huge body of research evidence, much of it is methodologically weak (Reynolds & Wheldall, 2007). There are very few randomised controlled studies of its efficacy. The WWC site has recently given it a positive report but this is based on only a handful of studies, having rejected many others and not always on a factually accurate basis. One of the scientific evaluations most widely cited in favour of Reading Recovery efficacy was, in fact, completed by the second author's research team (Center et al., 1995), as described above, and found that Reading Recovery was probably effective for only one in three students who experienced it and tended to be effective for those students who were least phonologically challenged. While the WWC site clearly regards it as Level 1, some might argue for a much lower level (see Reynolds & Wheldall [2007] for a recent review of studies on Reading Recovery).

A program such as Jolly Phonics, however, being based on the most up-to-date research evidence, making clear conceptual sense, and with further supportive evidence for specific program efficacy based on randomised controlled trials (see, for example, Stuart, 1999) would be a strong candidate for a Level 1 grading. We might then consider programs that are predicated on sound scientific research evidence but for which there is little or no specific evidence for efficacy. There are a number of seemingly sound phonics-based programs that would fit into this category. They would tend to be located at Level 3.

But what about programs such as the widely promoted treatment offered by the Dore Centres, formerly known as DDAT, which claim to achieve extraordinary results in the treatment of dyslexia? (Claims are also made for the success of the method in treating attention deficit hyperactivity disorder and even Asperger's syndrome.) The treatment proposed by the Dore Centres appears to be essentially predicated on a widely discredited model, the perceptual motor program. Such programs have a long and far from illustrious history in special education. In spite of considerable accumulated evidence that such programs are ineffective, they resurface every decade or so under a different name or guise (Stephenson et al., 2007). Moreover, the two scientific studies of Dore's efficacy published in a refereed scientific journal (*Dyslexia*) have subsequently been severely challenged and criticised by numerous reading researchers and *Nature*, arguably the most influential science journal in the world, has seen fit to publish a cautionary editorial (*Nature*

Neuroscience, 2006). Dore, then, would probably currently locate at Level 4, or even 5, on the proposed scale.

Basing Educational Policy and Practice on Science

The phrase, 'the elephant in the room', has become almost a media cliché of late to signify what is manifestly obvious to all but unvoiced. We might, then, ask what is the guilty secret, obvious but unspoken, of ineffective school education? It is not, as many pundits from teachers' associations would have us believe, too little funding, too few resources or too few staff. The elephant in the classroom is that most classroom instruction is simply not good enough. We would argue that the reasons for this are at least threefold.

First, as we have seen, teachers do not appear to be operating from an empirical database of scientific fact. They are not trained to do so and their subsequent professional reading of educational research is minimal at best, as discussed earlier in this article. Rather, or perhaps as well as, being encouraged to be 'reflective practitioners', we should also require teachers to become more like 'scientist-practitioners'. Rather than reflecting on what is 'just good teaching', their teaching should be informed by the findings from rigorous scientific research which has successfully identified the critical components of effective instruction and classroom practice. Research in teaching children with special needs, for example, has yielded a set of teaching skills and strategies that have been shown to be consistently effective. Some contemporary special educators have learned to be what are sometimes called 'scientist-practitioners' or 'data-based teachers'. Data-based teachers are, firstly, teachers who are sensitive to research findings on effective teaching methods. Rather than being guided by fashion and hype or the opinions of others, they look at the research findings. They evaluate the data and make their judgements on the basis of empirical evidence. They also collect data themselves so that their own teaching is guided by data. They systematically monitor the performance of their students and change what they do on the basis of this information. They also monitor their own teaching performance. On the basis of this continual monitoring they make educational decisions and change their practice accordingly. But they are few and far between.

The second reason why classroom teaching is not good enough is that the zeitgeist ('spirit of the times') in many, if not most, education faculties in which Australians are taught to be teachers would seem to be based on an avowedly constructivist approach to education (see, for example, Standing Committee on Employment, Workplace Relations and Education, 2007). Much of the apparent research literature promoting constructivist pedagogy, however, appears to be more descriptive or exhortative than evidence-based. Apps and Carter (2006), for example, refer to a pilot study they conducted in which they searched the ERIC database from 1982 to 1999 for the terms *constructivism* and *discovery learning*, and also the term *direct instruction* as a comparison reference point. According to Apps

and Carter, the search revealed that while discovery learning produced 1871 hits and constructivism 1170 hits, direct instruction produced fewer than half as many, 409 hits. More important, however, was their subsequent more detailed analysis of the abstracts of the first 50 and the last 50 articles within each category. As Apps and Carter comment, their results:

> illustrated the increase of constructivist literature and revealed a tendency for this literature to be primarily of a non-empirical nature. For example, 51% of articles addressing direct instruction were empirical and examined student learning outcomes, compared with 2% of articles addressing discovery learning and 4% addressing constructivism. (p. 8)

(In a subsequent study specifically addressing constructivist approaches to special education, they examined all 114 peer reviewed articles up to October 2004 on this topic revealed by searches of both ERIC and PsychINFO and found that only 6 [5.3%] were experimental in nature.) These findings suggest that there is considerably more empirical work to be done before the evidence can match the rhetoric advocating constructivist approaches to teaching. Interestingly, a recent Federal Government senate committee examining quality of school education (Standing Committee on Employment, Workplace Relations and Education, 2007) expressed a considerable degree of reservation regarding the influence of constructivism on education. And yet this is the guiding philosophy routinely taught to those aspiring to become teachers in many of our faculties of education.

The third reason why classroom teaching is simply not good enough is that government education agencies ignore the scientific evidence that is available, even when they have commissioned the research themselves. It would seem unremarkable, for example, to suggest that the model of reading instruction provided in our schools should be informed by scientifically validated best practice. Over the past 30 years, we have seen the growth of a huge body of scientific research literature internationally, illuminating both how reading works and how it should best be taught. Reading instruction that includes serious attention to phonics has been shown repeatedly and conclusively to be the most effective method of teaching reading (see, for example, Coltheart & Prior, 2007). Note that we argue 'includes serious attention to phonics', not 'places exclusive emphasis upon phonics'. (Reading education is too important to trivialise with extremist political posturing.)

This is not a matter of opinion; it is a matter of established, replicated, verifiable, scientific fact. And yet, when the 'Nelson Report' of the 'National Inquiry into the Teaching of Literacy' (NITL) was released in December 2005 (Department of Education, Science and Training, 2005), strongly advocating an explicit, systematic phonics-based approach to reading instruction in our schools, it was all but ignored. Since the Nelson Report was released, there has been little done of appreciable significance to implement its findings. More seriously, what has been done has been paying little more than lip-service to the Report's major recommendations.

Nowhere is this perhaps more manifest than in the implementation of the second phase of the Reading Assistance Voucher (RAV) scheme. While it would be inappropriate to address this issue in detail here (see Wheldall [2007] for further details), it provides a pertinent example of how subsequent government behaviour fails to follow the recommendations of its own committees of inquiry. In brief, the Department of Education, Science and Training invited tenders for the production of a Reading Assistance Kit (RAK) for use by tutors in assisting low-progress readers that fully complied with the recommendation of the (NITL) Report. The set of materials produced for the RAV scheme was subsequently severely criticised by the Chair of NITL, Dr Ken Rowe (of the Australian Council for Educational Research), who, according to the *Australian* newspaper of April 5 2007, said: 'the tutorial resources failed to teach basic skills required to read, such as the relationship between sounds and letters … Their lack of alignment with the recommendations (of the Inquiry's report) is extraordinary'. This would appear to be clear evidence of either the unwillingness or the complete inability of government to allow educational policy to be determined by the best available scientific evidence.

In both the UK and the United States, however, there have been serious attempts to tie increased educational funding to redress the problems of poor literacy standards with requirements that the funding be spent on programs of demonstrable efficacy. The Reading First initiative within the No Child Left Behind legislation in the United States clearly required that the additional federal funds on offer to state educational facilities were to be spent exclusively on reading initiatives which were in accord with the available scientific evidence on reading instruction, notwithstanding the previously documented problems with the WWC initiative. Educational funding in Australia has not traditionally been tied to demonstrations of efficacy. The only accountability requirement is financial. But the solution to providing effective education is not solely dependent on funding; more money does not necessarily mean better programs. Effective education is tied to government commitment to supporting what actually *works* rather than what is *fashionable*.

Conclusion

In many cases it can be argued education, including much special education, exhibits the conspicuous external trappings of science but the core values of pseudoscience. While we can lament the failure of governments to implement evidence-based practices in education, it is probably not reasonable to expect such action when we fail to comprehensively embrace such values as a profession. While currently deeply flawed, attempts to address the evidence base in education, such as the WWC, do represent a step in the right direction. With the present poverty of gold standard evidence and consequent flaws in decision making by the WWC, an alternative approach examining levels of evidence, such as the one outlined in the current article, may represent a viable temporary solution.

References

American Cancer Society. (n.d.). Vitamin C. Retrieved June 12, 2007, from http://www.cancer.org/docroot/ETO/content/ETO_5_3X_Vitamin_C.asp?sitearea=ETO.

Apps, M., & Carter, M. (2006). When all is said and done, more is said than done: Research examining constructivist instruction for students with special needs. *Australasian Journal of Special Education*, *30*, 107–125.

Barrett, S., & Jarvis, W. T. (Eds.) (1993). *The health robbers: A close look at quackery in America*. Buffalo, NY: Prometheus.

Boardman, A. G., Arguelles, M. E., Vaughn, S., Hughes, M. T., & Klingner, J. (2005). Special education teachers' views of research-based practices. *Journal of Special Education*, *39*, 168–180.

Carnine, D. (2000). Why education experts resist effective practices (and what it would take to make education more like medicine). Retrieved June 12, 2007, from http://www.edexcellence.net/doc/carnine.pdf.

Carroll, R. T. (2005). The skeptic's dictionary: Blondlot and N-rays. Retrieved 12 June, 2007, from http://skepdic.com/blondlot.html.

Center, Y., Wheldall, K., Freeman, L., Outhred, L., & McNaught, M. (1995). An evaluation of Reading Recovery. *Reading Research Quarterly*, *30*(2), 240–263.

Coltheart, M., & Prior, M. (2007). *Learning to read in Australia.* Occasional Paper of the Academy of the Social Sciences in Australia, 1/2007 (Policy Paper #6).

Department of Education, Science and Training. (2005). *Teaching reading*. Canberra: Department of Education, Science and Training.

Forness, S. R. (2001). Special education and related services: What have we learned from meta-analysis? *Exceptionality*, *9*, 185–197.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*, 149–164.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*, 165–179.

Huizenga, J. R. (1993). *Cold fusion: The scientific fiasco of the century*. Oxford: Oxford University Press.

Iverson, S., & Tunmer, W. E. (1993). Phonological processing skills and the Reading Recovery program. *Journal of Educational Psychology*, *85*, 112–126.

Jacobson, J. W., Foxx, R. M., & Mulick, J. A. (2005). *Controversial therapies for developmental disabilities: Fad, fashion, and science in professional practice*. Mahwah, NJ: Lawrence Erlbaum.

Landrum, T. J., Cook, B. G., Tankersley, M., & Fitzgerald, S. (2002). Teacher perceptions of the trustworthiness, usability, and accessibility of information from different sources. *Remedial and Special Education*, *23*, 42–48.

Nature Neuroscience. (2006). Editorial. *Nature Neuroscience*, *10*, 135.

NSW Institute of Teachers. (n.d.). Continuing professional development policy. Retrieved October 15, 2007, from http://www.nswteachers.nsw.edu.au/IgnitionSuite/uploads/docs/Continuing%20Professional%20Development%20Policy.pdf.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. D., Thompson, B., & Harris, K. (2004). *Quality indicators for research in special education and guidelines for evidence-based practices: Executive summary.* Retrieved 15 June, 2007, from http://education.uoregon.edu/grantmatters/pdf/DR/Exec_Summary.pdf.

Reynolds, M., & Wheldall, K. (2007). Reading Recovery 20 years down the track: Looking forward, looking back. *International Journal of Disability, Development and Education*, *54*, 199–223.

Rudland, N., & Kemp, C. (2004). The professional reading habits of teachers: Implications for student learning. *Australasian Journal of Special Education*, *28*, 4–17.

Sagan, C. (1997). *The demon haunted world: Science as a candle in the dark*. London: Headline.

Sasso, G. M. (2001). The retreat from inquiry and knowledge in special education. *Journal of Special Education*, *34*, 178–193.

Seethaler, P. M., & Fuchs, L. S. (2005). A drop in the bucket: Randomized controlled trials testing reading and math interventions. *Learning Disabilities Research & Practice*, *20*, 98–102.

Shanahan, T., & Barr, R. (1995). Reading recovery: An independent evaluation of the effects of an early instructional intervention for at-risk learners. *Reading Research Quarterly*, *30*, 958–996.

Shane, H. C. (1994). *Facilitated communication: The clinical and social phenomenon*. San Diego, CA: Singular Press.

Shermer, M. (1997). *Why people believe weird things: Pseudoscience, superstition, and other confusions of our time*. New York: Henry Holt.

Slavin, R. E. (2003). A reader's guide to scientifically based research. *Educational Leadership*, *60*(5), 12–16. Retrieved October 15, 2007, from http://www.ascd.org/portal/site/ascd/menuitem.459dee008f99653fb85516f762108a0c/.

Standing Committee on Employment, Workplace Relations and Education. (2007). *Quality of school education*. Canberra, Australia: Commonwealth of Australia.

Stephenson, J. (2004). A teacher's guide to controversial practices. *Special Education Perspectives*, *13*, 66–74.

Stephenson, J., Carter, M., & Wheldall, K. (2007). Still jumping on the balance beam: Continued use of perceptual motor programs in Australian schools. *Australian Journal of Education*, *51*(1), 6–18.

Stephenson, J., & Wheldall, K. (2008). Miracles take a little longer: Science, commercialisation, cures and the Dore program. *Australasian Journal of Special Education*, *32*(1), 67–82.

Stuart, M. (1999). Getting ready for reading: Early phoneme awareness and phonics teaching improves reading and spelling in inner-city second language learners. *British Journal of Educational Psychology*, *69*, 587–605.

Viadero, D. (2006, September 27). 'One stop' research shop seen as slow to yield views that educators can use. *Education Week*, pp. 8–9.

Victorian Institute for Teaching. (n.d.). Renewal of Registration. Retrieved October 15, 2007, from http://www.vit.vic.edu.au/files/documents/1133_renewal-info-brochure-2007.pdf.

What Works Clearinghouse. (2006a). Evidence standards for reviewing studies. Retrieved June 15, 2007, from http://www.whatworks.ed.gov/reviewprocess/study_standards_final.pdf.

What Works Clearinghouse. (2006b). WWC intervention report: Arthur. Retrieved July 31, 2007, from http://www.whatworks.ed.gov/PDF/Intervention/WWC_Arthur_091406.pdf.

What Works Clearinghouse. (2006c). WWC intervention report: Arthur appendix. Retrieved July 31, 2007, from http://www.whatworks.ed.gov/PDF/Intervention/techappendix10_259.pdf.

What Works Clearinghouse. (n.d.a). Extent of evidence categorization. Retrieved July 31, 2007, from http://www.whatworks.ed.gov/reviewprocess/extent_evidence.pdf.

What Works Clearinghouse. (n.d.b). Intervention rating scheme. Retrieved July 31, 2007, from http://www.whatworks.ed.gov/reviewprocess/rating_scheme.pdf.

Wheldall, K. (2007). Turning a blind eye to Nelson. *Bulletin of Learning Difficulties Australia*, *39*(1), 1–2.