© The Policy Press • 2005 • ISSN 0547 4378

debate

Is the evidence-based practice movement doing more good than harm? Reflections on Iain Chalmers' case for research-based policy making and practice

Martyn Hammersley

English Powerful voices are currently insisting that policy and practice must be based on research evidence, and that social science inquiry should be reformed in order to serve this need more effectively. An influential figure in the evidence-based practice movement is Sir Iain Chalmers, previously director of the UK Cochrane Centre. Taking evidence-based medicine as his model, he presents the task of research as to determine which policies and practices work. This is to be achieved through the use of randomised controlled trials and systematic reviews of their results. In this article, some of the central assumptions of his case are assessed.

Français Un courant d'opinion puissant insiste actuellement sur le fait que les politiques générales et les pratiques doivent être basées sur les preuves de la recherche et que les investigations de science sociale devraient être réformées, afin d'aller plus efficacement dans cette direction. Sir Iain Chalmers, directeur précédent du Centre Cochrane du Royaume-Uni, est un personnage influent du mouvement de pratique basée sur les preuves. En prenant la médicine basée sur les preuves comme modèle, il présente la mission de recherche en tant que détermination des politiques et pratiques qui fonctionnent. Ceci doit être accompli à travers l'utilisation d'essais contrôlés randomisés et de l'examen systématique de leurs résultats. Dans cet article, certaines des suppositions centrales de son cas sont évaluées.

Español Las voces actuales poderosas insisten en que la política y la práctica deben basarse en una evidencia de investigación, y que las preguntas de ciencias sociales deberían reformarse para servir a esta necesidad de manera más efectiva. Una figura de gran influencia en el movimiento de la práctica basada en la evidencia es el señor Iain Chalmers, el anterior director del centro Cochrane en el Reino Unido. Tomando la medicina basada en la evidencia como su modelo, él representa la tarea de investigación para determinar qué políticas y qué prácticas funcionan. Esto es para conseguir a través del uso de pruebas aleatorias controladas y de análisis sistemáticos sus resultados. En este artículo se evalúan algunas de las asunciones centrales de su caso.

Key words	evidence-based practice • social research • randomised controlled trials • systematic reviews
Key dates	final submission 12 April 2004 • acceptance 12 May 2004

The idea that policymaking and practice should be 'evidence based' has become widely accepted. And, at face value, it is not difficult to understand why. Who would want policy or practice *not* to be based on evidence? Yet, of course, while this is how the issue is often presented, it is not what is at stake[1]. The evidence-based practice movement argues that policymaking and practice should be based on *research* evidence presented in the form of *systematic reviews*, in other words syntheses of the findings from all relevant studies meeting some threshold of methodological rigour. Furthermore, the notion of rigour applied here – both to primary research and to the task of reviewing – assumes a conception of research methodology that is broadly positivist in character (Hammersley, 2001; Loughlin, 2003)[2]. As a result, the evidence-based practice movement raises some fundamental issues for many social scientists: about the proper role of research evidence, as against other sources of information, including personal experience, in policymaking and professional practice; about the role of evidence of all kinds versus the role of judgment in decision making; about how the relevance of research evidence should be defined in reviewing the literature, and whether or not it is necessary to carry out exhaustive searches; about how methodological soundness is to be determined, and how important it is; about whether and how the findings from different studies can be 'synthesised' in a worthwhile manner; and, finally, about the kinds of contribution that research can make to policymaking and practice.

The evidence-based practice movement began in the field of health, but its influence has now extended out across other fields, including education, criminology, and social work (Trinder, 2000; Thomas and Pring, 2004). The impact has varied considerably, but there has been resistance in these areas, just as there was (and continues to be) in medicine. In this article, I want to look at the case made for research as supporting evidence-based practice in a recent article by Iain Chalmers (2003), one of its most influential advocates in medicine, and someone who also champions its extension to other fields. In the course of his article, 'Trying to do more good than harm in policy and practice: the role of rigorous, transparent, up-to-date evaluations', he responds to the arguments of critics, including my discussions of systematic review (Hammersley, 2001, 2002a).

Chalmers (2003, p 22) summarises a leading theme of the evidence-based practice movement in the opening sentence of the abstract to his article:

> Because professionals sometimes do more harm than good when they intervene in the lives of other people, their policies and practices should be informed by rigorous, transparent, up-to-date evaluations.

In the field of medicine, this demand was stimulated by recognition that many clinical judgments were not based on up-to-date research evidence, that the use of randomised controlled trials and syntheses of their findings could provide a means for assessing these judgments, and, above all, by the discovery that some of the standard treatments employed by clinicians did not appear to produce improvements in patients, a few even worsening their conditions. In the course of extending this argument from the core areas of drug testing and the assessment of surgical techniques to other aspects of health service provision, and beyond these areas to further kinds

of professional practice, there have been efforts to incorporate the findings of other sorts of research, even qualitative inquiry, into systematic reviews, or to synthesise these findings in ways that are more sensitive to the character of the original studies (see Dixon-Woods et al, 2004). Nevertheless, a particular model of scientific practice remains central to the evidence-based practice movement; one that emphasises what is seen as the key role of 'transparent' procedures in ensuring rigour.

Let me outline, to begin with, those parts of Chalmers' argument where I think there is, or ought to be, broad agreement. He is surely correct that professional practitioners in all fields sometimes do things which are not beneficial, and occasionally do harm. Furthermore, research has an important role to play in providing information for policymaking and practice. It can frequently offer knowledge about how policies have actually been implemented, about variation in this across contexts, and about the *effects* of policies or practices, intended and unintended, foreseen and unforeseen. We might add, too, that it can tell us how and why particular policies and practices become influential at particular places and times. At the same time, there is a strong tendency for all manner of ideas to be presented as if they were research evidence when they are not the product of research and/or are not very reliable. These ideas may come from governments, from commercial organisations, from inspection regimes or auditing agencies, from think tanks of one sort or another, from researchers making claims that exceed their competence, and from other sources as well. Within the public sphere, and particularly on the part of the media, there is insufficient attention to how any information reported or used was produced, and in particular to likely sources of error involved in it. In fact, even within the realm of research, in my view, there is a great deal of over-claiming. This stems, in part, from demands that researchers demonstrate the practical value and interest of their work. Finally, I agree with Chalmers that the results of research should be presented to lay audiences through reviews of the available literature, rather than the findings of individual studies being offered as reliable information. It is common for researchers to believe that the findings from their own investigations should be acted on by policy makers and practitioners, and indeed they are encouraged by funding bodies to disseminate their findings to 'users' so that these can 'impact' on practice. Yet the findings of individual studies are at best usually only steps toward sound knowledge, rather than being reliable in themselves. Reviews are an essential bridge between the worlds of research and those of policy making and practice (Hammersley, 2002b, ch 7).

Despite these areas of agreement, I have serious doubts about key arguments used by Chalmers, and many of these are the stock-in-trade of advocates of evidence-based policymaking and practice (see, for example, Oakley 2000). My doubts can be summarised in relation to his leading theme:

- Can we ever prevent professionals sometimes doing more harm than good?
- Can we determine with great certainty via research alone *whether* they are doing harm or good?
- Does evaluating policies and practices by means of research always lead to more good than harm?

All of these questions are begged from the very start of Chalmers' article, and yet they need to be addressed[3].

These general doubts arise from more specific concerns. Central here is the sharp distinction between practitioner *opinion* and scientific research *evidence* that is built into Chalmers' argument and into other presentations of the rationale for evidence-based practice. He points out that while clinicians are often convinced about the value of the treatments they employ, surveys show that application of these treatments varies across practitioners, so that individual certainty is at odds with what he calls 'collective uncertainty'; and he portrays research evidence as able to 'adjudicate' among these conflicting opinions by documenting what does and does not work (Chalmers, 2003, pp 23-4). However, I will argue that there are several reasons why this sharp distinction is not justifiable, and why we should not treat research evidence as able to play the role of adjudicator. Establishing this will require me to look at both sides of the practitioner opinion/scientific evidence divide.

The role of research in practice

Chalmers believes that professional practice should be *based on* research evidence. However, as critics have pointed out, and some advocates of evidence-based practice recognise, this idea is potentially misleading: practice is necessarily a matter of judgement, in which information from various sources (not just research) must be combined[4]. Moreover, no evidence is infallible, so we would expect policy makers and practitioners to assess critically the claims made in research reports (even in systematic reviews), and in doing so they will necessarily draw on their own experience and background knowledge. In addition, there are problems concerning how one applies research evidence about aggregates to particular cases, and about how one weighs the implications of such evidence against information from other sources where the two conflict (see Byrne, 2004). In his account, Chalmers effectively assumes that research evidence should always prevail (in providing adjudication); and while this does not seem to be the line taken by all advocates of evidence-based medicine (see, for example, Sackett et al, 1997, p 72; Davies, 1999, p 111), the problems that face practitioners in seeking to operate in a research-informed fashion have not been sufficiently addressed (Oswald and Bateman, 2000). It is important to emphasise that research evidence cannot serve as a court of appeal for judging competing conceptions of best practice, in the way that Chalmers and most other advocates of evidence-based practice believe. This is because research cannot supply all the information that practitioners require in order to engage in good practice, and because (as already noted) research findings must always be interpreted and are never free from potential error. Moreover, the various sources of knowledge on which practical judgment relies are often not commensurable; they cannot be 'weighed' in terms of the same scale. A more complex process is required; for example, knowledge from personal experience and from new research evidence must each be evaluated in its own terms, and then combined in some way that takes account of their distinctive characteristics as sources of knowledge[5].

It is also important to remember that there may be significant variation across different types of professional practice in the role that judgment plays and in the

contribution of sources of information other than research[6]. For example, evidence from research may be capable of playing a stronger role in some areas of medicine than others, and in medicine as against social work and education. Putting the point the other way round, the role of practical judgment may be greater in some fields than it is in others simply because of the nature of the problems professionals have to deal with and the circumstances in which they must act. So, in any assessment of the role of research, the distinctive requirements of professional practice in the area concerned need to be taken into account[7].

Chalmers is preoccupied with bias and complacency on the part of practitioners: their reliance on unrepresentative personal experience and on outdated knowledge, and the tendency to treat current practices as if their value were beyond doubt. While such dangers are always present, however, what he is pointing to here is simply bad practice: practitioners' reliance on personal experience and on what they were taught in the past need not be uncritical, and their judgments need not be dogmatic. Moreover, while it is difficult to know how widespread such bad practice is, even if it is common we cannot assume that requiring practitioners to become familiar with the latest research evidence, and to base their decisions on this, will improve matters. Sound judgment would only be produced if this evidence, and its implications for practical decisions in the context of other information, were evaluated wisely. And, in fact, the injunction to treat research evidence as if it can adjudicate over what is best practice is unlikely to encourage such wisdom, since it underplays the proper role of judgment in practical decision making (see Eraut, 1994, ch 3; Bilson and White, 2004).

The nature of research evidence

Turning to the other side of the contrast, Chalmers treats research as producing practical recommendations whose likely validity is much greater than that of those based on professional experience[8]. However, there are reasons to doubt this assumption. A first point concerns the fallibility of all research findings. It is important to recognise that like all other forms of human practice research itself necessarily relies on judgment and interpretation: it can never be *governed*, but only *guided*, by methodological rules. Furthermore, whatever efforts are made to prevent bias from extraneous factors, whether these are the beliefs of researchers or pressures from sponsors, we can never *ensure* that no distortion has occurred. Finally, not only are the findings of research fallible, like those from all other sources of knowledge, but there are also distinctive threats to validity involved in studying human social relations. In what follows, I will look at various validity threats surrounding research evidence, both in relation to randomised controlled trials and to systematic reviews.

Chalmers places emphasis on the capacity of randomised controlled trials to provide clear evidence about 'what works', although he does not completely rule out other kinds of research evidence. And it is true that such trials are a valuable research strategy in some fields. Nevertheless, carrying them out and interpreting their findings are not quite as straightforward as Chalmers implies in this article[9]. Moreover, the problems, in both respects, vary across areas of investigation.

Even in well-run drug trials, there are threats to validity, especially 'external validity',

because the samples studied are not representative of the relevant wider population, and therefore cannot always tell us what works for whom, or about the incidence of side effects[10]. There is also the problem that 'what works' is always bound up with some assumptions about the mechanisms by which it works. And these assumptions are part of what is being tested in the trial[11]. Furthermore, while Chalmers insists that blind allocation is not a defining feature of randomised controlled trials, whether a trial is blind – in relation to those administering treatments and/or those receiving them – can be a significant factor affecting the likely internal validity of the findings. And such blinding is not always feasible, even in testing the effectiveness of drug treatments[12].

The problems are more serious when we move away from drug trials. There, at least in the experimental context, the treatment is usually relatively specific and can be standardised, and the outcomes can sometimes be measured with little likely error (for example, in terms of survival rates, although even here there are problems where *cause* of death needs to be identified). By contrast, in many other fields, treatments cannot be standardised in the same way and/or outcomes cannot be measured very reliably. For example, assessing the effects of a particular pedagogical strategy on children's learning is much more difficult (because the 'treatment' is not fixed in character and the effects are hard to measure accurately), with the result that the findings are very much more uncertain in their validity[13]. Furthermore, the problems faced here are not simply practical difficulties in implementing randomised controlled trials; they are to do with the very nature of what is being studied. For instance, the behaviour of schoolteachers cannot easily be standardised because a requirement for effectiveness in the job is adaptation to circumstances, notably to the distinctive and changing characteristics of particular cohorts of children. Furthermore, how children respond to the use of a specific pedagogical strategy depends to some extent on how they interpret the teacher's behaviour, both whether they understand what is being required of them and what attitude they take towards this. In more specific terms, a teacher's actions always carry potential messages for pupils about his or her expectations about them; and what expectations pupils ascribe to the teacher can influence their learning (Rogers, 1982).

Another way of putting this point is that, built into randomised controlled trial methodology is a rather simple conception of causality, of how a treatment generates outcomes (Byrne, 2004; see also Cook and Payne, 2002). This model may be closely approximated in some fields, and it could serve as a useful guide in other areas too, but we need to remember that there are fields where it does not seem to apply and may be seriously misleading. This problem can be illustrated by the fact that there has been a great deal of educational research over the past 100 years concerned with identifying the features of effective teaching; in other words, 'what works' in education. Much of this has used experimental or quasi-experimental method, but the results do not suggest (to say the least) that simple causal relations can be found, even though what has been discovered is by no means worthless (Dunkin and Biddle, 1974; Gage, 1985, 1994; Chambers, 1991, 1992; Glass, 1994; Floden, 2001; Hamilton and McWilliam, 2001). Those, like Chalmers, who wish to extend randomised controlled trials to areas outside medicine need to take account of what has already been attempted in those fields, and what might be learned from this, rather than

assuming that the randomised controlled trial is a magic bullet that can be applied anywhere to provide an accurate and precise evaluation of 'what works'.

Let me turn, next, to Chalmers' advocacy of systematic review. Here, again, he takes much for granted and largely ignores the problems that have been identified. He presents systematic review as simply an application of scientific method to the task of reviewing; and he interprets this method as relying on procedural objectivity. Yet this ignores what we have learned about the nature of natural science, and much of what has been discovered in doing social science: the practice of inquiry cannot be rendered 'transparent' in terms of explicit rules and the validity of the results thereby ensured; and while, up to a point, following guidelines may improve the quality of scientific work, beyond that point it may cause damage because methodological rules come to be applied unthinkingly – in other words, without appropriate judgment[14].

There are also questions about the purpose of 'systematic reviewing' and how this relates to the various functions that more traditional reviews can be designed to serve. The literature on systematic reviewing tends to promote a single conception of the review process: that it is concerned with pooling the data or findings from multiple studies in order to maximise the accuracy and precision of conclusions about which policies or practices work. Yet this is not the only useful way of synthesising research findings; and synthesis of this kind is not the only (or even an essential) task of reviews. Equally important in conventional forms of reviewing is bringing studies together that are complementary, in the sense of providing knowledge about different aspects of the same phenomenon. The purpose of the review process also has implications for whether exhaustive searches for relevant studies are necessary. If one's aim is to represent what is recognised as well-established knowledge in a field, it may not be necessary to carry out such searches. Instead, the focus should be on those studies, or sets of studies, that have come to be accepted as sound and productive by researchers. Nor are exhaustive searches always required if the aim is the development and testing of theoretical ideas through drawing on the secondary literature (on which see Pawson, 2002). Moreover, even where exhaustive searching is appropriate, it must be remembered that there are costs involved: the time and other resources spent on searches cannot be devoted to reading and evaluating particular studies, synthesising findings, and writing the review (Hammersley, 2002a).

Another area of criticism concerns the approach that systematic reviewers often take to assessing the validity of studies relevant to the focus of their review. The problems here are several. First, there is still a strong tendency – evidenced in this article by Chalmers – to assume that randomised controlled trials offer the best evidence, and qualitative case studies the worst; whereas, even in abstract terms, what they offer is different types of data involving varying threats to validity. Second, there is a tendency to assume that a standard set of criteria should be applied in evaluating studies, whereas it may sometimes be necessary to use different criteria according to the function a study is being used to serve within a review. After all, the purpose of reviews, often, is not simply to identify what is sound knowledge but also to indicate what might be fruitful future lines of inquiry. These may be suggested by studies that would be ruled out on the basis of most sets of validity criteria. Third, and more fundamentally, built into much advocacy of systematic reviews is the idea

that the likely validity of findings can be judged on the basis of the research design adopted in the study. What is meant by 'research design' here, for example, is whether any physical (or statistical) control was imposed on extraneous factors in coming to judgments about what treatments produce what outcomes. This is certainly an important issue, but it is not a simple one; and it is not the only important consideration in judging the likely validity of research findings. For one thing, not all studies are concerned with explanation or theory testing. Some are descriptive, and here there may be no need for controlled comparisons (although sometimes there is). A second point is that the notion of physical control is not straightforward in the context of studying human behaviour. In applying physical controls here, one usually increases reactivity and thereby endangers ecological validity. Another way of putting the same point is that, strictly speaking, there can be no such thing as physical control in experiments on human beings because people respond in terms of signs and interpretations rather than on the basis of physical reflexes (Rosnow, 1981). There is also the problem of what people's reaction to variation in a single factor can tell us about how they will behave in relation to that factor when it is experienced in contexts where other relevant considerations are not controlled[15].

The implication of these points is not that physical control is unimportant or not worthwhile, but rather that it does not always maximise the validity of findings and nor is it always necessary to produce valid findings. Furthermore, assessing the validity of research findings always depends on judging their plausibility in terms of what is currently taken as known and in terms of likely threats to validity in their production. This is not something that can be governed by some standardised and fully explicit procedure. Guidelines that remind us of what must be taken into account will be valuable so long as they are not treated as routine checklists; but they cannot eradicate the judgement that is involved in the process. Nor should we think of this judgment as inevitably a source of bias, any more than methodical rigour is necessarily a source of validity. Judgements do not simply reflect the characteristics of the judge: they can involve skilled and knowledgeable assessment of what is likely to be true. They may be wrong, and they may be biased, but they need not be – and they are essential in making any methodological assessment.

For all these reasons, relating to the nature both of professional practice and of research, it is important to question the sharp distinction that Chalmers, and others, draw between practitioner opinion and research evidence. There is a failure to recognise the implications of the fallibility of scientific evidence, that reliable evidence can derive from other sources besides research, and that using any evidence requires judgment, both about its validity and about what its implications for practice might be in particular contexts. In insisting on this sharp contrast, advocates of evidence-based practice mirror the position taken by some postmodernists. Where one side places excessive emphasis on the role of scientific knowledge as a corrective to subjective judgment, the other rejects science as ideological and oppressive in favour of reliance on what it regards as necessarily arbitrary judgment[16]. Despite their opposition to one another, both positions advocate a sharp distinction between knowledge and judgment that is not defensible. More sensible is an approach that

recognises a spectrum of kinds of evidence that are not always commensurable, and that necessarily depend on practical judgment if they are to be used wisely.

Chalmers' response to criticism

It is striking that Chalmers does not engage seriously with the specific points made by those who criticise the privileging of randomised controlled trials and systematic review; and in this he is in line with some other proponents of evidence-based practice (for example, Gough and Elbourne, 2002; Oakley, 2003). Instead, the criticisms are dismissed as 'political' (Oakley), 'ideological' (Gough and Elbourne), or as 'polemic' (Chalmers, 2003, p 28). While complaining that critics set up straw arguments, Chalmers does the same, as when he claims (2003, p 30) that:

> those who reject randomization are implying they are sufficiently knowledgeable about the complexities of influences in the social world that they know how to take account of all potentially confounding factors of prognostic importance, including those they have not measured, when comparing groups to estimate intervention effects.

Apparently, as far as Chalmers is concerned, to suggest that there are limits to the value of randomised controlled trials implies a rejection of randomisation and a belief in one's own omniscience; whereas one might suggest instead that it simply indicates an appropriate level of humility and caution. In much the same way, a writer who raises ethical concerns is accused of "bizarre ethical analyses" (2003, p 30)[17]. Critics are also accused of "ignoring evidence" and failing to "confront reality" (2003, p 36). In this way, rather than engaging with the arguments of critics, Chalmers simply treats what they say as symptomatic of their alleged ignorance, stupidity, or ulterior motives.

In his response to my own work, Chalmers directly misrepresents what I write. He claims that I reject "the notion that bias 'can and must be minimised', because this is 'assumed to maximise the chances of producing valid conclusions'". And on this basis declares that I have a "cavalier lack of concern about bias in reviews" (Chalmers, 2003, p 26). Not only is the view he ascribes to me incoherent, as a result of highly selective quotation, but the first part of it is the opposite of what I believe. My argument was that there is a positivist assumption embodied in systematic review to the effect that subjectivity is necessarily a source of bias, and so must be minimised through proceduralising the review process. I suggested that this assumption is false, that subjectivity is not always a source of bias, any more than proceduralising research guarantees valid results. I certainly did not deny that bias can and must be minimised in order to maximise the chances of producing valid conclusions. Indeed, minimising bias is, for me, the core of scientific inquiry (see Hammersley and Gomm, 2000), and is an essential element in the reviewing process. What I do deny is that minimising subjectivity (where that term is not defined as equivalent to error) is necessary or indeed desirable. In relation to this piece of misrepresentation, Chalmers' own term of abuse – 'cavalier' – seems self-referential[18].

There is, then, little attempt on Chalmers' part to understand the criticisms made of the case he and others have presented for evidence-based practice. While he

recognises that the critics may have different views from him about the nature and functions of research, he dismisses these views as "ideology parading as intellectual inquiry" (Mosteller and Boruch, 2002, p 2, quoted in Chalmers, 2003, p 26). This is very much what one would expect from advocates seeking to promote the interests of a social movement[19]. It is not conducive, however, to discovering what is and is not sound in the notion of evidence-based practice.

Conclusion

There are some important arguments put forward by Chalmers and other supporters of the evidence-based practice movement. Nevertheless, their conclusions are seriously misleading because they are based on too sharp a distinction between practitioner opinion and research evidence. This leads them to make excessive claims for the role that research can play in guiding policymaking and practice.

From the very start of his article, Chalmers is keen to secure the ethical high ground, insisting on the need for research to evaluate practice in order to prevent harm. He refers (2003, p 32) to the "human and financial costs arising from failure to perform systematic, up-to-date reviews of randomized controlled trials of health care". However, the grounds for his assumption that research can *adjudicate* over what is best practice are weak. Indeed, one might charge him with the offence of exaggerating the capacity of research to resolve practical problems, and of presenting research as capable of providing evidence whose validity is more certain than it often is. Moreover, if policymakers and practitioners are encouraged to give the findings of research more weight than they deserve, this could result in undesirable outcomes: policies or treatments may not be used when they would have been of value, or be treated as more reliable than they actually are. The point is that there can be collateral costs to carrying out randomised controlled trials and systematic reviews; they can themselves cause harm (a possibility that Chalmers neglects). My argument here is not that the fallibility of research undermines its value; it does not. There is, however, a need to balance the one-sided account that Chalmers provides of its benefits. There is probably room for improvement in all forms of practice, and research evidence may often be able to play a key role in bringing this about, but it cannot guarantee to improve rather than to worsen the situation. We need to face that disturbing fact rather than ignoring it.

There is also the problem that Chalmers and others portray a certain kind of research as being able to answer policy and practical problems directly, as if there could be technical solutions, based on facts, to practical problems that necessarily involve value judgments. In some areas of medicine, this may not matter, because there are no competing values or goals. One of the examples that Chalmers uses is a case in point: the recommendation that babies should be put to sleep on their backs probably follows virtually automatically from evidence to suggest that this significantly reduces the dangers of sudden infant death. By contrast, however, in evaluating various methods of teaching reading, for example, what is aimed at by these methods may not be the same, so that the differences arise at least partly from variation in educational and other values. In these circumstances, portraying research as showing 'what works' can serve as an ideological device that closes down proper

discussion about the relative weight that should be given to different educational goals. While research can provide evidence about the consequences of various policies, on its own it cannot tell us what is the best thing to do, either in general terms or in particular cases.

Of course, these points are not new. Many of them have long been recognised by those engaged in the task of evaluating social interventions. The field of evaluation research was founded in the US in the 1960s as a result of concern about the effectiveness and efficiency of such initiatives. Initially, researchers in this area attempted to apply experimental method to the task of evaluation, but they soon discovered serious problems with this initial orientation, and the result was a subsequent diversification in approach (Shadish et al, 1991; Pawson and Tilley, 1997, ch 1). While some of the arguments used against early attempts at evaluation and in favour of later approaches may be spurious, there is no justification for returning to the initial naive faith in experimental method; a more subtle assessment of its strengths and weaknesses is required. In other words, anyone advocating randomised controlled trials in the field of social policy needs to take account of what can be learned from past experience of carrying out evaluations in that field. Failure to do so is unlikely to enhance good practice.

As I noted, a disturbing feature of Chalmers' article, shared with some other advocacy of evidence-based practice, is his apparent unwillingness to engage in serious discussion about the issues: his response to critics is to dismiss them as ignorant, cavalier, ideological, and so on, sometimes on the basis of outright misrepresentation. This orientation is, of course, characteristic of political movements that are concerned with sustaining and building their position in the world, and in some respects the evidence-based practice movement matches this type. Towards the end of his article, Chalmers (2003, p 38) argues that "uncertainty and humility among policymakers, practitioners and researchers" are desirable, and complains that these attitudes are "in short supply". He is surely right about this, yet he only sees these attitudes as preconditions "for wider endorsement" of the approach he is advocating. In other words, humility and uncertainty are required on the part of others so that they are in the right frame of mind to accept what he himself claims to know with certainty. We might suggest, instead, that these virtues are appropriate for us all, given the difficulties involved in reaching sound conclusions about the effects and desirability of particular interventions. Only on that basis will we be able to make sober assessments of how research can best contribute to policymaking and practice.

Notes

[1] On the rhetorical ploy involved here, see Tanenbaum (2003) and Hammersley (2002a). For background to the development of the evidence-based practice movement, see Trinder (2000). For an outline of the relationship between social science and the notion of evidence-based policymaking, see Young et al (2002).

[2] What Loughlin (2003), following Goodman (2002), calls 'Star Trek epistemology'.

[3] For a parallel assessment of the assumptions underlying evidence-based practice, specifically in the field of medicine, see Norman (1999).

[4] Chalmers himself seems to recognise this here and there, but it is at odds with the main force of his argument, in terms of which the essential role of practitioners' judgments is played down.

[5] For discussions of this problem from rather different directions, see Dunne (1993, chs 9, 10, especially pp 46-7) and Djulbegovic et al (2000).

[6] Such variation was recognised long ago by Aristotle; see Dunne (1993, ch 8, especially pp 253-61).

[7] For an elaboration of this argument in relation to evidence-based practice in education, see Hammersley (2002b, ch 1).

[8] Elsewhere, he takes what seems to be a very different position when he declares that "a leap of faith will always be required to make causal inferences about the effects of health care" (Chalmers, 1995, p 1315); however, he goes on to indicate that, for him, 'reliable evidence' will 'usually mean evidence derived from systematic reviews of carefully controlled evaluative research'. Surely what is required is neither reliance on procedures that purportedly maximise the validity of conclusions nor leaps of faith but, rather, reasonable judgments taking relevant evidence properly into account.

[9] On the difficult practicalities surrounding trials, see Marks (1997) and Gueron (2002).

[10] For discussion of an RCT involving serious threats to external validity, including those arising from recruitment problems, see Dehue (2002, pp 89-91; 2004). There are also some fundamental problems with the very distinction between internal and external validity which undercut claims that properly conducted experiments maximise validity; see Hammersley (1991).

[11] This relates to a deeper problem with Chalmers' position: he seems to assume a regularity theory of causation, whereby trials can identify fixed relationships between treatments and outcomes that operate outside the experimentally controlled situation across cases and over time. This assumes that there are relatively stable, closed systems of relationships in all the fields where the evidence-based approach is to be applied. This is probably not an assumption in which we ought to have great confidence; see Byrne (2004).

[12] For an indication of the sorts of bias that can result from the absence of blinding, see Dehue (2002, p 88). This is an issue that has long been given attention in the literature on psychological methodology; see Rosnow (1981). It is also worth pointing out that randomisation was not central to experimental methodology until the 1930s, and its value has been challenged recurrently. Interestingly, its first use, albeit not in the form of random allocation of subjects but of treatments to the same subject, seems to have been in Charles Sanders Peirce's psychophysical experiments; see Hacking (1988).

[13] Chalmers simply ignores these issues in claiming that randomised controlled trials have demonstrated the superiority of phonics-based methods of teaching children to read (Chalmers, 2003, p 23). The same problems arise in many other fields: see Kippax's (2003) discussion of the experimental evaluation of sexual health interventions, the import of which is also ignored by Chalmers. Also relevant is Wolff's (2000) account of the problems involved in using RCTs in the field of mental health care. Prideaux (2002) makes similar points regarding the evaluation of problem-based learning in medical education, but once again his arguments are dismissed by Chalmers.

[14] Elsewhere, I have used the work of Polanyi to develop this argument (Hammersley, 2002a). Much the same point is central to the work of Kuhn; see Bird (2000).

[15] This is an argument developed by Egon Brunswik; see Hammond and Stewart (2001).

[16] The work of Lyotard exhibits postmodernist celebration of the arbitrariness of judgment. On this, see Drolet (1994).

[17] Graebsch's (2000) argument is about a proposal legally to allow RCTs without properly informed consent on the part of those subjected to the treatments. She argues that this is against the principle of equal treatment and against the ethical principle never to use another person solely as a means, a principle that was enshrined in the German legal system as a result of scientific abuses under the Third Reich. Whatever one's judgment about her argument, which is in large part about the legality of RCTs in relation to criminal justice in Germany, there is nothing 'bizarre' about it.

[18] Chalmers (2003, p 26) goes on to accuse me of "unfamiliarity with the field of research synthesis". His argument here also relies on a misinterpretation, as well as being patently false.

[19] On the political sociology of the evidence-based practice movement, see Marks (1997); Traynor (2000); Dehue (2002); Tanenbaum (2003); and Hammersley (2004).

References

Bilson, A. and White, S. (2004) 'Limits of governance: interrogating the tacit dimensions of clinical practice', in A. Gray and S. Harrison (eds) *Governing medicine: Theory and practice*, Maidenhead: Open University Press.

Bird, A. (2000) *Thomas Kuhn*, Princeton (NJ): Princeton University Press.

Byrne, D.S. (2004) 'Evidence-based: what constitutes valid evidence?', in A. Gray and S. Harrison (eds) *Governing medicine: Theory and practice*, Maidenhead: Open University Press.

Chalmers, I. (1995) 'What do I want from health research and researchers when I am a patient?', *British Medical Journal*, vol 310, no 6990, pp 1315-18.

Chalmers, I. (2003) 'Trying to do more good than harm in policy and practice: the role of rigorous, transparent, up-to-date evaluations', *Annals of the American Academy of Political and Social Science*, vol 589, pp 22-40.

Chambers, J.H. (1991) 'The difference between the abstract concepts of science and the general concepts of empirical educational research', *Journal of Educational Thought*, vol 25, no 1, pp 41-9.

Chambers, J.H. (1992) *Empiricist research on teaching: A philosophical and practical critique of its scientific pretensions*, Boston, MA: Kluwer Academic.

Cook, T.D. and Payne, M.R. (2002) 'Objecting to the objections to using random assignment in educational research', in F. Mosteller and R. Boruch (eds) *Evidence matters: Randomized trials in education research*, Washington, DC: Brookings Institution.

Davies, P. (1999) 'What is evidence-based education?', *British Journal of Educational Studies*, vol 47, no 2, pp 108-21.

Dehue, T. (2002) 'A Dutch treat: randomized controlled experimentation and the case of heroin-maintenance in the Netherlands', *History of the Human Sciences*, vol 15, no 2, pp 75-98.

Dehue, T. (2004) 'Historiography taking issue: analyzing an experiment with heroin abusers', *Journal of the History of the Behavioral Sciences*, vol 40, no 3, pp 247-64.

Dixon-Woods, M., Agarwal, S., Young, B., Jones, D. and Sutton, A. (2004) *Integrative approaches to qualitative and quantitative evidence*, London: Health Development Agency (www.hda-online.org.uk/documents/integrative_approaches.pdf).

Djulbegovic, B., Morris, L. and Lyman, G. (2000) 'Evidentiary challenges to evidence-based medicine', *Journal of Evaluation in Clinical Practice*, vol 6, no 2, pp 99-109.

Drolet, P.M. (1994) 'The wild and the sublime: Lyotard's post-modern politics', *Political Studies*, XLII, pp 259-73.

Dunkin, M.J. and Biddle, B.J. (1974) *The study of teaching*, New York, NY: Holt, Rinehart and Winston.

Dunne, J. (1993) *Back to the rough ground: 'Phronesis' and 'techne' in modern philosophy and in Aristotle*, Notre Dame, IN: University of Notre Dame Press.

Eraut, M. (1994) *Developing professional knowledge and competence*, London, Falmer.

Floden, R.E. (2001) 'Research on effects of teaching', in V. Richardson (ed) *Handbook of research on teaching* (4th edn), Washington, DC: American Educational Research Association.

Gage, N.L. (1985) *Hard gains in the soft sciences: The case of pedagogy*, Bloomington, IN: Phi Delta Kappa.

Gage, N.L. (1994) 'The scientific status of the behavioural sciences: the case of research on teaching', *Teaching and Teacher Education*, vol 10, no 5, pp 565-77.

Glass, G.V. (1994) 'Review of Chambers' empiricist research on teaching', *Journal of Educational Thought*, vol 28, no 2, pp 127-30.

Goodman, K. (2002) *Ethics and evidence-based medicine*, Cambridge: Cambridge University Press.

Gough, D. and Elbourne, D. (2002) 'Systematic research synthesis to inform policy, practice and democratic debate', *Social Policy and Society*, vol 1, no 3, pp 225-36.

Graebsch, C. (2000) 'Legal issues of randomized experiments on sanctioning', *Crime and Delinquency*, vol 46, no 2, pp 271-82.

Gueron, J.M. (2002) 'The politics of random assignment: implementing studies and affecting policy', in F. Mosteller and R. Boruch (eds) *Evidence matters: Randomized trials in education research*, Washington, DC: Brookings Institution.

Hacking, I. (1988) 'Telepathy: origins of randomization in experimental design', *Isis*, vol 79, no 298, pp 427-51

Hamilton, D. and McWilliam, E. (2001) 'Ex-centric voices that frame research on teaching', in V. Richardson (ed) *Handbook of research on teaching* (4th edn), Washington, DC: American Educational Research Association.

Hammersley, M. (1991) 'A note on Campbell's distinction between internal and external validity', *Quantity and Quality*, vol 25, no 4, pp 381-7.

Hammersley, M. (1995) *The politics of social research*, London: Sage Publications.

Hammersley, M. (2001) 'On "systematic" reviews of research literatures: a "narrative" response to Evans and Benefield', *British Educational Research Journal*, vol 27, no 5, pp 543-54.

Hammersley, M. (2002a) 'Systematic or unsystematic: Is that the question? Some reflections on the science, art, and politics of reviewing research evidence', Paper presented to the Public Health Evidence Steering Group of the Health Development Agency (www.hda.nhs.uk/evidence/sys_unsys_phesg_hammersley.html).

Hammersley, M. (2002b) *Educational research, policymaking and practice*, London: Paul Chapman.

Hammersley, M. (2004) 'Some questions about evidence-based practice in education', in G. Thomas and R. Pring (eds) *Evidence-based practice in education*, Maidenhead: Open University Press.

Hammersley, M. and Gomm, R. (2000) 'Bias in social research', in M. Hammersley *Taking sides in social research*, London: Routledge.

Hammond, K.R. and Stewart, T.R. (2001) 'Introduction', in K.R. Hammond, and T.R. Stewart (eds) *The essential Brunswik: Beginnings, explications, applications*, New York, NY: Oxford University Press.

Kippax, S. (2003) 'Sexual health interventions are unsuitable for experimental evaluation', in J.M. Stephenson, J. Imrie and C. Bonell (eds) *Effective sexual health interventions: Issues in experimental evaluation*, Oxford: Oxford University Press.

Loughlin, M. (2003) 'Ethics and evidence-based medicine: fallibility and responsibility in clinical science [Kenneth Goodman, essay review]', *Journal of Evaluation in Clinical Practice*, vol 9, no 2, pp 141-4.

Marks, H.M. (1997) *The progress of experiment: Science and therapeutic reform in the United States 1990-1990*, New York, NY: Cambridge University Press.

Mosteller, F. and Boruch, R. (eds) (2002) *Evidence matters: Randomized trials in education research*, Washington, DC: Brookings Institution Press.

Norman, G.R. (1999) 'Examining the assumptions of evidence-based medicine', *Journal of Evaluation in Clinical Practice*, vol 5, no 2, pp 139-47.

Oakley, A. (2000) *Experiments in knowing: Gender and method in the social sciences*, Cambridge: Polity Press.

Oakley, A. (2003) 'Research evidence, knowledge management and educational practice: early lessons from a systematic approach', *London Review of Education*, vol 1, no 1, pp 21-33.

Oswald, N. and Bateman, H. (2000) 'Treating individuals according to evidence: why do primary care practitioners do what they do?', *Journal of Evaluation in Clinical Practice*, vol 6, no 2, pp 139-48.

Pawson, R. (2002) 'Evidence-based policy: the promise of "realist synthesis"', *Evaluation*, vol 8, no3, pp 340-58.

Pawson, R. and Tilley, N. (1997) *Realistic evaluation*, London: Sage Publications.

Prideaux, D. (2002) 'Researching the outcomes of educational interventions: a matter of design, *British Medical Journal*, vol 324, pp 126-7.

Rogers, C. (1982) *A social psychology of schooling: The expectancy process*, London: Routledge and Kegan Paul.

Rosnow, R.L. (1981) *Paradigms in transition: The methodology of social inquiry*, New York, NY: Oxford University Press.

Sackett, D.L., Richardson, W.S., Rosenberg, W. and Haynes, R.B. (1997) *Evidence-based medicine: How to practice and teach EBM*, New York, NY: Churchill Livingstone.

Shadish, W.R., Cook, T.D., and Leviton, L.C. (1991) *Foundations of program evaluation: Theories of practice*, Newbury Park, CA: Sage Publications.

Tanenbaum, S. (2003) 'Evidence-based practice in mental health: practical weaknesses meet political strengths', *Journal of Evaluation in Clinical Practice*, vol 9, no 2, pp 287-301.

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

Traynor, M. (2000) 'Purity, conversion and the evidence-based movements', *Health*, vol 4, no 2, pp 139-58.

Trinder, L. with Reynolds, S. (eds) (2000) *Evidence-based practice: A critical appraisal*, Oxford: Blackwell Science.

Wolff, N. (2000) 'Using randomized controlled trials to evaluate socially complex services: problems, challenges and recommendations', *Journal of Mental Health Policy and Economics*, vol 3, no 2, pp 97-109.

Young, K., Ashby, D., Boaz, A. and Grayson, L. (2002) 'Social science and the evidence-based policy movement', *Social Policy and Society*, vol 1, no 3, pp 215-24.

The response to this article can be seen in *Evidence & Policy*, vol 1, no 2, pp 227-242.