Social Science and Evidence-based Everything: the case of education

ANN OAKLEY, Social Science Research Unit, University of London Institute of Education, London, UK

ABSTRACT Recent moves in academic and policy circles to strengthen the social science research evidence base have raised questions about the quality and status of educational research. They have suggested a need for systematic research synthesis, for greater accessibility of sound educational research evidence, and greater respect for the perspectives of the different stakeholders in the educational research process. This paper looks at the background to ‘the evidence movement’, and discusses a particular government-funded initiative designed to take forward the challenge of systematic reviews of educational research. It considers some of the ways in which this activity poses challenges for social science methodology.

A Brief History of Evidence-based Everything

There is a long history of people’s dissatisfaction with expert and non-systematic approaches to evidence, mostly in medicine but also in other areas (Oakley, 1998). Early in the twentieth century the science of research synthesis as a remedy for this dissatisfaction began to emerge with Karl Pearson’s classic study in 1904 on the effectiveness of typhoid vaccine (Pearson, 1904), and with other work in education, agriculture and physics. Many of the tools of research synthesis were developed by American social scientists from the 1960s on. Notable in this effort were the educational psychologist Gene Glass and his wife, psychotherapist Mary Lee Smith, who computed 833 effect sizes derived from 375 studies obtained from over 1000 citations in order to find out whether or not psychotherapy worked (Smith & Glass, 1977). Their conclusion that psychotherapy does work made the psychotherapists happier than their colleagues in the social work field, who around the same time had to deal with a discouraging review about casework which suggested that troubled clients left to their own devices did no worse, and sometimes distinctly better, than those who were ‘helped’ by social workers (Fischer, 1973).

Current Challenges

Although research synthesis began in social science, today’s evidence movement in education and social policy more generally is very much driven by the example of what can be achieved with research synthesis in health care. The Cochrane Collaboration (http://www.cochrane.org), an international network of health care researchers, has established 50 different collaborative review groups which undertake pro-
grammes of systematic reviews, contributing the results to an electronic database of research evidence, the Cochrane Library. This currently holds 1297 completed reviews, with another 1013 in protocol stage.

The example of the Cochrane Collaboration has made professionals and policymakers in other disciplines think hard about the parallels and differences between health care and other forms of professional intervention in people’s lives. As the President of the Royal Statistical Society said in 1996: ‘We are, through the media, as ordinary citizens, confronted daily with controversy and debate across a whole spectrum of public policy issues. But typically we have no access to any form of a systematic ‘evidence base’— and therefore no means of participating in the debate in a mature and informed manner’ (Smith, 1996, pp. 369–70).

These criticisms have been extended to education in the form of a number of arguments about the unscientific, non-cumulative, uncollaborative and inaccessible nature of much educational research (Hargreaves, 1996; Hillage et al., 1998; McIntyre & McIntyre, 1999; OECD, 2001; Tooley & Darby, 1998). There have been many responses to these criticisms. Is the objective of basing policy and practice on evidence simply a government plot? Is it an attempt to restrict academic freedom by imposing standard criteria for research quality—much like the attempts to standardise tests for educational attainment and the quality of teachers’ work? Is it a wildly mistaken move to impose the ‘positivism’ of the medical model on a social world that cannot possibly be expected to meet the same requirements? (Atkinson, 2000; Elliott, 2001; Hammersley, 2001). Of course, each of these arguments is analytically distinct, but their coalescence at a historical moment when the notion of ‘evidence’ is coming at us from all quarters suggests a rhetorical linkage between them.

One way to consider these criticisms is to take a closer look at some of the work that is actually going on under the general heading of ‘evidence-based/informed education’.

The EPPI Centre and Research Synthesis

In 2000 the DfES funded a programme of systematic reviews of educational research supported by the Evidence for Policy and Practice Information and Coordinating (EPPI) Centre at the Social Science Research Unit, University of London Institute of Education. The DfES initiative at the EPPI Centre is currently funded until 2005 and covers schooling from 0–19. Its aims are: to set up and coordinate groups of researchers and others who want to do systematic reviews in education; to provide an infrastructure of training and support in how to do these; to develop standardised systems for collecting and classifying bibliographic references and for undertaking, holding and updating reviews; to coordinate methodological work; and to develop ways of working with different groups of users. The vision at the end of this work is REEL—The Research Evidence in Education Library. This is an electronic database of completed reviews of educational research evidence— and the hope is that it will be available in every school and educational institution in the country, and that it will be user-friendly and relevant to the needs and interests of a large range of stakeholders in the education process.

The nearest parallel to REEL is the Cochrane Library. However, we in education have the opportunity to benefit from the experiences of the Cochrane Collaboration in sponsoring systematic reviews of health care. One example of what we hope will
Evidence-based Everything

not happen in educational research synthesis can be illustrated by a recent enquiry about what is known about the best way to treat the epidemic of head lice in schoolchildren. Most Internet search engines will locate a large number of different sites relating to this topic, and many of these give contradictory advice. The Cochrane Library does have a recently updated review of interventions to treat head lice (Dodd, 2002), and this is superior to most Internet offerings in providing a systematic account of where the information included in it has been taken from, and how it has been quality assessed. The head lice treatment review is 45 pages long, examines 70 studies, dismisses 67 of them as being of unsound quality and makes detailed recommendations for future research needs. But the review is full of technical language and, having read the review as a user of health care, one is not much nearer knowing what one ought to do to treat a child with head lice than one was at the beginning.

If the notion of ‘evidence’ is to mean anything other than the intellectual property of elite groups, the accessibility of both the process and the results of research synthesis to a range of users must be an integral value. The EPPI Centre work is based on the following principles: that different user groups need access to high quality review of research; that collaborative developments are needed in systematic review methodology; and that such moves towards a firmer evidence base for social science and policy will get nowhere unless they follow principles of openness and equal participation of all stakeholders. Like all good things, this sounds wonderful. It is unlikely to be easy.

During our first two years, ten groups of researchers and practitioners have been through a formal registration system to become review groups (RGS). The groups are in the following areas: English teaching; assessment and learning; school leadership; gender and education; post-compulsory education; inclusive education; early years; thinking skills; modern languages; and the impact of continuous professional development on classroom teaching and learning. The methods for facilitating the review work of these groups were piloted in a systematic review of strategies to support pupils with emotional and behavioural difficulties in mainstream primary classrooms conducted by two researchers at the Institute of Education and the National Foundation for Educational Research (Evans & Benefield, 2001). Two years into the initiative and members of the EPPI education team are beginning to reflect on their experiences in carrying out the first reviews (see Torgerson et al., in press, for an example). These reflections identify both successes and challenges, and will form an important learning experience for future research synthesis in education. They point to particular gaps in the methodology of research synthesis, among which the lack of agreed quality criteria for establishing the validity and reliability of ‘qualitative’ research is probably the most critical. Although there has been some progress in this direction, for example in the field of health promotion reviews (Harden et al., 2001; Shepherd et al., 2001), it is as yet unclear how this is likely to translate into the related field of educational inquiry.

The registration process for RGS includes peer refereeing of plans and protocols. Each review undertaken by a review group requires a detailed review protocol which defines the review question, inclusion and exclusion criteria and strategies for searching the literature. All protocols are placed on the EPPI Centre website for open comment. Most reviews are done in two stages: a mapping stage, in which relevant literature is captured and systematically keyworded to provide a descriptive account of the research effort in that particular area; and an in-depth review stage, in which
a subset of the literature found is examined and interrogated in more detail and data extracted from primary studies. An important feature of the reviewing software developed by the EPPI Centre is that it allows data to be entered on a range of study designs. These data are then analysed together for individual reviews, but they can be combined differently to answer other review questions; reviews can also be updated by adding data from new primary studies.

The function of the groups is to produce systematic reviews which are part of the REEL electronic database of educational research reviews. The RGs decide the questions to be tackled in the reviews, and what kinds of studies they wish to include. The review process follows a clear set of stages. It begins with setting the review question and developing the rest of the protocol, continues through searching for studies, working out what literature to include at title and abstract stage, keywording the studies found and entering them into a bibliographic database, deciding which studies to look at in depth, extracting data from these, putting the data into a synthesis database, undertaking some quality assessment to identify reliable studies, and finally synthesising the evidence and writing the review report.

**Systematic Reviews**

The above list includes accepted stages for a systematic review (Davies, 2000). A recent paper by Mark Petticrew (2001) at the Social and Public Health Sciences Unit in Glasgow listed eight myths about systematic reviews: that they are just like ordinary reviews, only bigger; that they involve statistical synthesis; that they have to be done by experts; that they can be done without proper information and library support; that they are a substitute for good quality primary research; that they are not relevant to the real world; they require a biomedical model with easily quantifiable outcomes; and only include randomised controlled trials.

However, and contrary to its mythical representation, a systematic review is simply a way of accessing research knowledge; in this sense it is a piece of research in its own right. Systematic reviews synthesise the results of primary research, use explicit and transparent methods, and are accountable, replicable and updateable. Although the history of both medicine and social fields such as education are full of sporadic attempts to develop systematic methods for finding out what we know, most research reviews traditionally use either an expert model—in which experts are asked to say what they think—or a non-systematic approach (or both). Most traditional literature reviews are discursive rampages through selected bits of literature the researcher happens to know about or can easily reach on his or her bookshelves at the time. Neither in this approach, or in the expert model are the conclusions of a review based on anything explicit. It is very hard to know what authors are drawing on, and why they conclude what they do (Jackson, 1980).

Table I is from a programme of health education review work we have been doing for a number of years. We found six reviews of older people and accident prevention which included a total of 137 studies. But only 33 studies were common to at least two reviews, only two were common to all six reviews, and only one study was treated consistently in all six reviews. The second example, shown in Table II, is from studies of anti-smoking education for young people: two so-called systematic reviews carried out at the same time; 27 studies included in both reviews, only three common to both. The last figure in Table II—70—is the number of studies we knew
Evidence-based Everything

TABLE I. Six reviews of older people and accident prevention

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total studies included</td>
<td>137</td>
</tr>
<tr>
<td>Common to at least two reviews</td>
<td>33</td>
</tr>
<tr>
<td>Common to all six reviews</td>
<td>2</td>
</tr>
<tr>
<td>Treated consistently in all reviews</td>
<td>1</td>
</tr>
</tbody>
</table>

Oliver et al., 1999.

TABLE II. Two reviews of smoking prevention programmes for young people

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total studies included</td>
<td>27</td>
</tr>
<tr>
<td>Common to both reviews</td>
<td>3</td>
</tr>
<tr>
<td>Available for review at the time</td>
<td>70</td>
</tr>
</tbody>
</table>

Oakley & Fullerton, 1995

from our systematically keyworded bibliographic database were actually available to be included in these reviews at the time they were done.

The consequence of this situation is that reviewers reach very different conclusions about the nature of research evidence. In the above two examples, looking at what studies had been included and how they were treated explained the different conclusions of the different reviews about the best ways to stop older people falling over and persuading young people not to smoke.

Challenges for Social Science of Evidence-based Everything

There are at least four issues ahead of social science in accommodating itself to the challenge of the evidence movement. These are to: revisit critically the question of the differences between medicine and other professions; find ways of reducing bias in policy and practice evaluation; develop methods for assessing the trustworthiness of qualitative research; and soften the polemic of ‘quantitative’ versus ‘qualitative’ methods.

(i) Medicine, etc.

In 1999, Diane Ravitch, a Research Professor in Education, flew back to New York from a trip to California and the following day developed pains in her left leg together with a strange shortness of breath. Her neighbour, who was a radiologist, recognised the signs of pulmonary embolism, and sent her to hospital. There Ravitch listened to the doctors discussing her condition, and she began to fantasise about what would happen if the doctors who were gathered round her bed discussing her treatment were to be replaced with education experts. The first thing that would happen would be the disappearance of the doctors’ certainty that she had a problem. The education experts would argue about whether anything was actually wrong with her at all—after all, illness is a socially constructed experience. Some would point out that attributing problems to people is simply to blame the victim. In Ravitch’s fantasy (or nightmare) the hospital administrator walks in at this point and announces
that the hospital has just received a large grant to treat people like her. Immediately the experts decide that her symptoms are real after all, but now they are unable to agree on what to do. Each has their favourite cure, and each cites different sets of research ‘findings’ to support their case. Ravitch, who did get better, because there is an established and effective treatment for her complaint, was left wondering why medicine and education should behave so differently (Ravitch, 1999).

Of course, it is part of the whole process of professionalisation that we should defend our territory and our professional claims to expertise, and this is an endemic reason why educationalists, social workers, criminologists and so on behave as though what they do is unique, and not open to question. But is it really?

(ii) Reducing Bias

People doing research synthesis in both health care and other areas have been impressed by the way in which well-designed experimental studies versus other kinds of studies seem to behave differently when it comes to research findings. Table III shows the relationship between research design and successful versus unsuccessful interventions in two reviews of crime prevention research. Experimental evaluations using random or matched control groups were considerably less likely to demonstrate effectiveness than those using other designs.

There are many examples of practices which were considered helpful on the basis of observational evidence, but which rigorous evaluations showed to be harmful. A more recent example from the criminology field is a review of prison tour programmes for preventing juvenile delinquency. Known as ‘scared straight’, these initiatives employed the principle that seeing people in prison would deter young people from committing crimes. They have enjoyed a great deal of popularity in the USA, despite only ‘anecdotal’ evidence of their success. A systematic review by Anthony Petrosino and colleagues (Petrosino et al., 2000) found that ‘scared straight’ programmes actually have the opposite effect from the one intended, making crime more likely.

These findings from research synthesis are worrying for fields such as education where so much of the ‘evidence’ is derived from small-scale qualitative research, depends heavily on practitioner judgements about the right thing to do, and/or is taken from poorly evaluated interventions.

| Table III. Crime prevention research reviews: research design and effectiveness claimed |
|-----------------------------------|---------------------------------|---------------------------------|------------------|------------------|
|                                   | ‘Successful’ intervention | ‘Unsuccessful’ intervention | Random control groups | 27% | 73% |
|                                   |                               |                               | Non-random control groups | 50% | 50% |
|                                   |                               |                               | No control groups | 69% | 31% |
| Logan 1972 (N = 100 studies).     |                               |                               | Random/matched control groups | 25% | 75% |
|                                   |                               |                               | Other control groups | 51% | 49% |
(iii) Critically Appraising Different Study Designs

The third challenge, referred to earlier, concerns how to appraise different study designs. Whatever kinds of study we work with in research synthesis, it is necessary to make a decision about whether these have anything reliable to say or not. This ‘quality’ screening is needed for all types of research design included in a review. But, while there are lots of different scales for scoring the quality of experimental research, with some degree of consensus, for example on the importance on having socially comparable intervention and control groups, sorting out trustworthy from untrustworthy qualitative studies is a much more murky business.

Table IV shows the ‘quantitative’ criteria for assessing quality proposed by four research groups. The criteria included items such as ‘study purpose clearly stated’, ‘adequate and appropriate final sample’ and ‘careful records of data’, and added up to 46 in total, of which 28 occurred in only one of the lists, ten in two, and six in three, with only two being common to all four lists—a clear description of the sample and how it was recruited, and an adequate description of how the findings were derived from the data. The alternative is a list, shown in Table V, of ‘qualitative’ criteria for assessing the trustworthiness of qualitative studies proposed by another four groups. There were 28 different criteria in these four lists, including items such as ‘persistent observation’, ‘understanding data within holistic contexts’ and ‘the privileging of subjective meaning’. Thirteen of the criteria appeared once and 11 twice; one—the collection of ‘thick’ data—appeared in three lists, but none were common to all four.

<table>
<thead>
<tr>
<th>Table IV. ‘Quantitative’ criteria proposed by four groups* for judging the trustworthiness of ‘qualitative’ studies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total number of criteria</strong></td>
</tr>
<tr>
<td><strong>Used in:</strong></td>
</tr>
<tr>
<td>one review</td>
</tr>
<tr>
<td>two reviews</td>
</tr>
<tr>
<td>three reviews</td>
</tr>
<tr>
<td>all four reviews</td>
</tr>
</tbody>
</table>

*Boulton et al., 1996; Cobb & Hagemaster, 1987; Mays & Pope, 1995; Medical Sociology Group, 1996.

<table>
<thead>
<tr>
<th>Table V. ‘Qualitative’ criteria proposed by four groups* for judging the trustworthiness of ‘qualitative’ studies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total number of criteria</strong></td>
</tr>
<tr>
<td><strong>Used in:</strong></td>
</tr>
<tr>
<td>one review</td>
</tr>
<tr>
<td>two reviews</td>
</tr>
<tr>
<td>three reviews</td>
</tr>
<tr>
<td>all four reviews</td>
</tr>
</tbody>
</table>

*Leininger, 1994; Lincoln & Guba, 1985; Muecke, 1994; Popay et al., 1998.
(iv) Dissolving the Methods War

A main danger ahead is that large areas of research, traditionally important in education, escape the evidence net either because no one can reach any consensus about how to sort out the reliable from the unreliable, or (which might be worse) because a new orthodoxy sets in according to which qualitative research is simply a world apart—nothing to do with evidence at all. The last few years have seen a rash of papers on this theme. One recent one by a ‘qualitative’ sociologist (Barbour, 2001) argued that none of the five ‘technical fixes’ most cited in the methods literature as rescuing qualitative research from its rigourlessness—purposive sampling, grounded theory, multiple coding, triangulation and respondent validation—are anything more than spurious ‘bumper stickers’ designed to boost academic credibility.

Conclusion

The DfES-funded initiative at the EPPI Centre referred to in this paper is one of three listed at the beginning of the Strategy Consultation Paper produced by the National Educational Research Forum in 2000 as addressing the requirement for a coherent approach to existing educational research (NERF, 2000). The other two are the ESRC initiative on evidence-informed policy, coordinated by Ken Young at Queen Mary and Westfield College in London (http://www.evidencenetwork.org), and the Campbell Collaboration (http://campbell.gse.upenn.edu), an international network of people doing systematic reviews in education, social work and criminal justice. These three initiatives have common themes: the need to find out what we already know; the desirability of holding this knowledge in an easily accessible and updateable format; and the requirement that research should be more policy-relevant than it has sometimes been in the past.

Evidence-based everything, which includes evidence-informed education, represents a paradigm shift in thinking about the relationship between academic research and real world policy and practice. What exactly it will achieve remains to be seen, but it is unlikely to be a short-lived fashion. It poses particular challenges both for educational research and for social science approaches to knowledge synthesis and management more generally. The early experience of the EPPI Centre initiative for supporting research synthesis work in education suggests that it is possible to develop collaborative, democratic and systematic structures for reviewing research evidence, which will help to open up the traditionally rather esoteric world of educational research to public scrutiny.

Correspondence: Ann Oakley, Social Science Research Unit, University of London Institute of Education, 18 Woburn Square, London WC1H ONR, UK; E-mail: a.oakley@ioe.ac.uk

REFERENCES


