CHAPTER 6

Using Evidence to Inform Practice in Science Teaching

The Promise, the Practice, and the Potential

Judith Bennett

Overview

This chapter considers the background against which evidence-based initiatives have been introduced into education and science education. Overviews of the findings of two contrasting reviews are presented, and the experience of conducting these is used to assess what the methods have to offer to science education policy and practice.

The Promise: The Use of Evidence in Educational Research

The last decade or so has seen much written about the use of evidence in education, set in the wider context of the need to make more use of evidence in informing decisions about policy and practice in a range of public service areas, including health, social welfare and education (see, e.g., Davies, 2000). One question that is frequently asked is, 'What works?'

Certainly, it would be difficult to argue against the use of evidence to inform educational interventions, and it is important to know something about the likely effects of an intervention. However, underlying the question 'What works?' are a number of other questions:

- What constitutes good evidence?
- How might such evidence be gathered?
- How might such evidence be used to inform curriculum interventions?

At a time when increased emphasis was being placed on the role of evidence in informing decision making, there also was considerable debate over the usefulness of educational research in providing such evidence. In the United Kingdom, the debate was launched by David Hargreaves (1996). Hargreaves was critical of much educational research, arguing that schools would be more effective if teaching became a research-based profession, and he blamed researchers for failing to make this happen. Hargreaves argued that little of worth had emerged from half a century of educational research:

> *Given the huge amounts of educational research conducted over the past fifty years or more, there are few areas which*

> *have yielded a corpus of research evidence regarded as scientifically sound and as a worthwhile resource to guide professional action [p. 2].*

He went on to pose the question:

> *...[J]ust how much research is there which (i) demonstrates conclusively that if teachers change their practice from x to y there will be a significant and enduring improvement in teaching and learning, and (ii) has developed an effective method of convincing teachers of the benefit of, and means to, changing from x to y? [p. 5].*

Hargreaves also accused researchers of producing inconclusive and contestable findings of little worth and demanded an end to

> *...the frankly second-rate education research which has not made a serious contribution to fundamental theory or knowledge; which is irrelevant to practice; which is uncoordinated with any preceding or follow up research; and which clutters up academic journals which virtually nobody reads [p. 7].*

These were very serious criticisms of educational research, questioning its purpose, its rigour, its quality and its relevance. Unsurprisingly, there were strong rebuttals from the education research community, leading to extensive discussion in the literature about the nature and purpose of educational research (e.g., Norris, 1990; Tooley & Darbey, 1998; Hillage et al., 1998; Davies et al., 2000; Evans & Benefield, 2001, Hammersley, 2001; Oakley, 2002; Vulliamy, 2004). Neither was the debate limited to the United Kingdom, because similar themes formed the bases of discussion in the United States (see, e.g., example, Shavelson & Towne, 2002; Slavin, 2002).

Underpinning the criticism of education research was the notion that it was 'unscientific' because it failed to draw on the experimental approaches of the natural sciences, thus failing to yield recommendations for practice that could be implemented with confidence. The solution was seen to lie in the undertaking of 'high-quality' research rather than basing decisions on 'poor-quality' research, current 'whims', tradition or professional wisdom. Such research would be more scientific in design and provide a much more rigorous means of testing any educational intervention to assess its effectiveness. Hargreaves was one of a number of people who encouraged the educational research community to look to the medical research model for procedures and practices which would allow much more definite conclusions to be reached about 'what works'.

At the heart of evidence-based medicine is the desire to ensure that a particular treatment offered to a patient is based on scientific evidence which suggests that the treatment is likely to be more effective than any alternative. The key features of evidence-based medicine are the randomised, controlled trial (RCT), in which people (patients in health-care situations) are randomly allocated to groups receiving different treatments, with the outcomes being subjected to tests of statistical significance. Included in the treatments may be an option of no treatment or the use of a placebo, where the patient is unaware that the treatment will have no effect. In medical research, systematic reviews are used to synthesise the findings of series of interventions, allowing knowledge to be built up cumulatively. Such reviews have been used in medical research for some years, emerging from the setting up of the Cochrane Collaboration in Oxford in 1993. The Cochrane Collaboration draws on the principles described by its founder, Professor Archibald Cochrane, then president of the Faculty of Community Medicine of the Royal Colleges of Physicians in the United Kingdom, in his very influential book *Effectiveness and Efficiency: Random Reflections on the Health Services* (Cochrane, 1972). The Cochrane Collaboration advocates the use of quantitative research studies based on experimental methods, supported by systematic reviews of the findings of studies, to generate evidence on which decisions can be made.

In response to the criticisms of educational research, initiatives were made in the early 2000s to introduce aspects of the evidence-based medical model into education. The first of these initiatives was the setting up of

the Campbell Collaboration in Philadelphia in the United States to review evidence from RCTs in education, criminology and other social sciences (see, e.g., Petrosino et al., 2000). Others include the establishing of the What Works Clearinghouse (What Works Clearinghouse, 2002; http://ies.ed.gov/ncee/wwc/) and the *Best Evidence Encyclopedia* (BEE; www.bestevidence.org) in the United States to review and summarise research for policy and practice and the Evidence for Policy and Practice Initiative Centre (EPPI-Centre; www.eppi.ioe.ac.uk) in the United Kingdom, with its associated electronic Research Evidence in Education Library (REEL; accessible from the EPPI-Centre homepage) to focus on systematic reviews of research evidence in key areas of education.

In summary, there has been considerable debate over the last fifteen years about the nature of educational research and a drive to improve its quality through adoption of a more scientific approach. The promise is that this will provide a much sounder evidence base to inform decision making. How does the practice live up to the promise?

What Are Systematic Reviews?

Reviews of research are undertaken for a variety of purposes. They may be an entity in themselves, such as an expert review paper in a journal, or they may form part of a bid for research funding or a section of a research thesis. Most research reviews are not currently 'systematic' in that they do not follow the procedures normally associated with systematic reviews.

Systematic review methods involve developing systematic search strategies for reports of research studies based on specific criteria, coding the studies against pre-specified and agreed characteristics, generating an overview or map of the area and then looking in detail at specific aspects of studies. As such, systematic reviews are undertaken with reference to a rigorous protocol for identifying and including research studies and synthesising the findings. Depending on the nature of the evidence reviewed, systematic reviews also may involve meta-analysis of the findings.

Conventional reviews, sometimes referred to as 'narrative reviews', differ from systematic reviews in several ways. The most obvious of these are that the authors of a narrative review have much more latitude in determining the search strategy, structure and scope of what is included in the review and the way in which findings are presented. Advocates of the systematic review (see, e.g., Cooper, 1998) see narrative reviews as having the potential for a high degree of personal preference and selectivity which lays them open to the criticisms of bias in reporting, discussion and emphasis, whereas systematic reviews provide a much sounder evidence base for decisions about policy and practice.

Systematic reviews have been proposed as a key early step that can be taken towards improving educational research. The idea is to review systematically the nature and quality of what already exists. This section looks at systematic review methods applied in the United Kingdom in the context of educational research.

The EPPI-Centre

The EPPI-Centre has been funded by the U.K. government for over a decade now to support systematic reviews of research evidence in areas concerned with schools and students up to eighteen years of age. The EPPI-Centre is based in the Social Science Research Unit at the Institute of Education in London and works in partnership with review groups located around the United Kingdom. The review groups for the three core curriculum subjects in England and Wales, English, science and mathematics, were established in the Department of Educational Studies at the University of York in the period 2001–3.

The provision of significant funding grants from the central government in the United Kingdom was one indicator of the level of interest in and aspirations for systematic review work in education. The work also has the support of the Organisation for

Economic Co-operation and Development. In reviewing educational research in the United Kingdom (OECD, 2002), the report commented that

> *The review team emphasises the value of the EPPI-Centre. Building up the methodologies for scientific reviews and exploiting the results for future research are the most important efforts currently needed for accumulating knowledge on educational research [p. 21].*

As such, there was considerable political impetus to fund a series of systematic reviews in a number of areas of education, including assessment and learning, citizenship education, English, mathematics, post-sixteen education, school leadership, science and thinking skills.

EPPI-Centre Review Methods

The EPPI-Centre aims to produce high-quality reviews of research findings that provide evidence accessible to a range of different user groups, including teachers, researchers and policy-makers. Each review is undertaken by a review group, which is a form of steering group whose membership includes policy-makers, teachers, school inspectors, academic researchers, teacher trainers and those involved in curriculum development work.

In essence, a systematic review carried out under EPPI-Centre methodology comprises seven main phases, as detailed in Table 6.1.

The in-depth review involves extracting a range of data from the study (termed 'data extraction') through answering over 100 questions. These enable the study to be evaluated in terms of the study aims and rationale, the research questions, the design methods, the methods used to collect and analyse the data, the results and conclusions and the quality of the reporting. These features are used to make an overall quality judgement about the study of high, medium or low, and these judgements underpin the final synthesis of the quality of the research evidence in answering the review question. The review reports generated, detailing the steps in the process and the substantive findings, are substantial documents, being some 20,000 words in length. A detailed account of the detail of the review methods may be found in Bennett et al. (2005), and details of the review tools may be found on the EPPI-Centre website (www.eppi.ioe.ac.uk).

Table 6.1. Phases of a Systematic Review

Phase	*Main Activity*
1	Identification of review research question and development of inclusion/exclusion criteria
2	Producing the review protocol
3	Searching and screening for potentially relevant research studies
4	Coding research studies against the inclusion/exclusion criteria
5	Producing an overview of research in the area – the systematic map
6	Conducting the in-depth review via data extraction of key features in the study
7	Production of the review report

The Practice: Examples of Systematic Reviews in Science Education

The Science Review Group at York has undertaken systematic reviews in three areas: the impact on students of the use of context-based and science-technology-society (STS) approaches to the teaching of science (Bennett et al., 2003, 2007; Lubben et al., 2004), the use and effects of small-group discussion work in science teaching (Bennett et al., 2004a, 2004b, 2010; Hogarth et al., 2004) and the impact of information and communication technology (ICT) on science teaching (Hogarth et al., 2006). In addition to the review reports, a number of journal articles on aspects of the reviews also have been published (Bennett et al., 2005, 2007, 2010). This chapter focuses on the findings of the reviews undertaken in the first two of these areas because the nature of the work means that they provide contrasting examples of reviews. The review of context-based/STS approaches encompasses a number of experimental studies, whilst the

review of the use of small-group discussions contains a much higher proportion of studies that are predominantly qualitative in nature. Thus the two reviews present a good opportunity to explore the review methodology in two different contexts.

The key aspects of each review are summarised below, followed by discussion of the findings and their implications for the use of evidence.

Use of Context-Based/Science-Technology-Society (STS) Approaches

The first set of reviews explored the effects of the use of context-based/STS approaches on student understanding of science and attitudes towards science. This area was seen as important because the use of such approaches in science teaching has been one of the more significant shifts in science teaching over the last two decades, particularly in the eleven- to eighteen-year age range. In the classroom, the use of such approaches might mean, for example, that students study medical diagnostic techniques in order to develop their understanding of electromagnetic radiation and atomic structure or look at a range of different fabrics and their uses to introduce ideas about materials and their properties. Advocates of context-based approaches believe that there are improvements in both understanding of science and attitudes towards science as a result of their use. Those who are less persuaded of the benefits believe that the use of context-based approaches means that students do not acquire a good grasp of underlying scientific ideas – in other words, understanding is adversely affected. The review wished to test these claims.

The review research question was: What evidence is there that teaching approaches that emphasise placing science in context and promote links between science, technology and society (STS) improve the understanding of science ideas and the attitudes towards science of eleven- to eighteen-year-old students?

Three reviews were conducted within this overall question. The first focused on attitudes and understanding and the second and third on gender and ability effects, respectively.

THE REVIEW FINDINGS

The searches yielded some 2,500 studies, of which sixty-one met the inclusion criteria for the review. The chief characteristics of the work are as follows: Fifty of the studies in the systematic map were carried out in the United States, the United Kingdom, the Netherlands and Canada. Forty-one studies were undertaken with students in the eleven- to sixteen-year age range and eighteen with students in the seventeen- to twenty-year age range. The emphasis on students in the eleven- to sixteen-year age range is likely to reflect the perception of this age group being very critical in terms of the decline in interest in science.

All sixty-one studies were evaluations (this was a criterion for inclusion in the review), with twenty-four employing experimental research designs, that is, using some form of control group. The remainder explored effects only on students experiencing the context-based/STS materials. Forty-four of the studies reported on attitudes and 41 on understanding. Of these, twenty-four reported on both these aspects.

Just over half the studies (thirty-five) focused on initiatives characterised as science. Where there was a single-subject focus, thirteen related to chemistry, ten to physics and three to biology. It is likely that the focus on chemistry and physics in the individual science disciplines reflects the motives for developing context-based materials in the first instance, with chemistry and physics being seen as subjects with a lower appeal than biology.

Test results, unsurprisingly, were the most commonly used measure in experimental studies and were used in almost two-thirds of the cases. Questionnaires and interviews featured more prominently in non-experimental studies. The most common outcome measures employed in studies were test results (twenty-seven studies), open questionnaires (twenty-seven studies), agree/disagree scales (twenty-one studies) and interviews (twenty studies).

The data extraction and judgements about quality indicated that seventeen studies were of medium quality or better, and the evidence presented below is based on these seventeen studies.

Making judgements about the quality of studies is not easy, particularly when they involve complex interventions, such as context-based/STS approaches. A set of criteria therefore was developed against which studies could be judged. These related to the focus of the study (understanding and/or attitude, with these as explicit independent variables), research design, the reliability and validity of the data-collection methods and tools (including the measures to assess understanding and/or attitude, the reliability and validity of data analysis, the sample size and the matching of control and experimental groups, the nature of the data collected (before and after intervention or post-intervention), the range of outcome measures and the extent to which the situation in which the data were collected was representative of normal classrooms.

EVIDENCE ON UNDERSTANDING OF SCIENCE IDEAS

The evidence on understanding of science ideas came from the findings of twelve studies, and seven of the twelve studies reported evidence that indicates that context-based/STS approaches develop a level of scientific understanding comparable to that of conventional courses. Four studies indicated that context-based/STS approaches lead to a better understanding of science ideas than conventional courses and one to poorer understanding.

The findings of two studies pointed to a particular issue related to the assessment of understanding when comparing context-based/STS courses with conventional courses which concerns the nature of the items used to provide measures of understanding. In the United Kingdom, in addition to external examinations at age eighteen+, the Royal Society of Chemistry (a prestigious scientific body) has a test bank of standard chemistry questions which it makes available to teachers to use if they so wish to assess their students' knowledge. One of the studies reported that students taking a context-based chemistry course got lower scores on this national test than students taking a conventional course. However, the same students did better than students taking the conventional course in their final external examinations. The overall standard of these final examinations is regulated by an external body, but students taking the context-based course sit examinations with context-based questions rather than more conventional questions. The standard assessment items in the national test more closely resemble questions on more conventional examination papers. One of the other studies reports a similar finding. The implication is that students on different types of courses are likely to perform better on assessment items that resemble the style of course they are following.

HOW BIG ARE THE EFFECTS?

There has been considerable emphasis on 'effect size' in recent research literature on evaluation studies as a means of quantifying the size of the difference in performance between two groups. Effect sizes tend to be described as 'small' if less than 0.2 and 'large' if greater than 0.4 (see, e.g., Cohen, 1969). Typically, educational interventions tend to have small effect sizes.

Of the four studies that report improved understanding, none reported effect sizes per se. Two studies presented sufficient statistical analysis for effect sizes to be calculated, and both had 'large' effects. Of these, one study had a particularly large effect. It is worth noting here that the instrument used to test levels of understanding was developed by the researchers themselves as part of an ongoing research and development programme on STS education, and the issues concerning style of assessment items mentioned earlier also may be of relevance here.

In summary, the review findings on understanding of science ideas appear to provide good evidence that context-based/STS approaches provide as good a development of understanding as more conventional

approaches. There is more limited evidence to suggest that understanding may be enhanced. There is some evidence to suggest that performance on assessment items is linked to the nature of the items used; that is, students following context-based/STS courses perform better on context-based questions than on more conventional questions.

THE EVIDENCE ON ATTITUDES TOWARDS SCIENCE AND SCHOOL SCIENCE

The evidence on attitudes towards school science and science comes from the findings of nine studies. By far the most common approach to gathering data on attitude was the use of inventories involving agreement/disagreement scales (Likert-type questionnaires). In all but one of the cases where these were employed, the instruments were developed by the researchers specifically for the study.

Seven of the nine studies reported evidence that indicates that context-based/STS approaches improve attitudes towards school science (or aspects of school science) and/or science more generally. Of these studies, three presented data that had been subjected to statistical analysis, and each indicated that the effects were statistically significant at the 0.05 level. In one case, there were sufficient data to calculate an effects size, and this was 0.67 – a large effect. (This evaluation tools used here had been designed by the developers of the intervention.) The remainder of the studies either employed simple descriptive statistics or gathered data for which statistical analysis was inappropriate.

One study reported evidence that indicates that context-based/STS approaches promote attitudes towards school science comparable to those promoted by conventional courses, and one study reported evidence that indicates that context-based/STS approaches have a negative effect on attitudes towards school science.

Three studies also collected data relating to subject choices beyond the compulsory period and/or career intentions because these are seen as important indicators of attitude towards the subject. Here, the evidence is mixed. Two studies reported increases in numbers electing to study science subjects, and one reported no change.

In summary, the review findings on attitudes towards school science and science appear to provide very good evidence that context-based/STS approaches foster more positive attitudes towards school science than conventional courses. There is more limited evidence to suggest context-based/STS approaches foster more positive attitudes towards science more generally than conventional courses and mixed (and limited) evidence on the impact of context-based/STS approaches on science subject choices in the post-compulsory period.

GENDER AND ABILITY EFFECTS

Five medium- to high-quality studies explored gender effects. Three of these suggested that gender differences in attitudes are reduced through the use of a context-based approach. Two studies suggested that girls in classes using a context-based/STS approach held more positive attitudes towards science than their female peers in classes using a conventional approach. There also was some evidence from one study to suggest girls following context-based/STS courses were more positive than their peers following conventional courses towards pursuing careers involving science, with results being significant at the 0.01 level. Taken together, these findings suggest that there is moderate evidence to indicate that context-based/STS approaches promote more positive attitudes towards science in both girls and boys and reduce the gender differences in attitudes.

Only one study, though of high quality, explored ability effects and reported that lower-ability students in classes using a context-based/STS approach developed a better conceptual understanding of science and held significantly more positive attitudes towards science than their lower-ability peers taking conventional courses. They also developed a better conceptual understanding of science and held significantly more positive attitudes towards science than their higher-ability peers in the same classes. All results were significant at the 0.01 level. With only one study reporting

on ability effects, it is not possible to reach any general conclusions.

Use of Small-Group Discussions in Science Teaching

The second set of reviews focused on the use of small-group discussions in science teaching. Many people involved in teaching and curriculum development in science believe that small-group discussions are an important tool in science teaching, motivating students and enhancing their learning in science. This is set in the context of wider aspirations for science teaching that include the promotion of scientific literacy (e.g., Millar and Osborne, 1998) and the use of constructivist approaches in science teaching (e.g., Driver and Bell, 1985). There is also a growing body of evidence that teachers would welcome support and guidance on running small-group discussions (see, e.g., Osborne et al., 2002; Levinson & Turner, 2001) because their introduction into science lessons challenges the established pedagogy of science teaching and places new demands on teachers. The review area had additional interest for the review group in that a high proportion of the research studies in the area were almost qualitative studies, thus testing a review methodology that seemed to group members to be more suited to quantitative experimental studies.

The review research question was: How are small-group discussions used in science teaching with students aged eleven to eighteen years, and what are their effects on students' understanding of science or attitude towards science?

Within this context, three reviews were conducted, focusing on the nature of small-group discussions in science, the effect of small-group discussions on students' understanding of evidence in science and the effect of different stimulus materials on understanding of evidence in science.

THE REVIEW FINDINGS

The searches yielded some 2,290 studies, of which ninety-four met the inclusion criteria for the review. Some of the chief characteristics of the work are as follows: Most of the studies report work that has taken place in the United States, the United Kingdom and Canada. The majority of work (sixty-nine studies) focused on small-group discussions in relation to student understanding. A substantial amount of the work (fifty-seven studies) also focused on the nature of the communication itself and collaborative skills associated with the discussion tasks given to student groups. Typical small-group discussions involved groups of three to four students emerging from friendship ties and lasted for at least thirty minutes. They also had individual sense making as their main aim (as opposed to, for example, leading to a tangible product such a as a poster or group presentation) and use prepared printed materials as the stimulus for discussion. Most of the work focused on biology or physics topics, with very little on chemistry topics. This appeared to reflect their use in addressing the difficulty of some physics topics, with small-group discussions being used as a means of students exploring their understanding of particular ideas, and the more issues-based of some biology topics, for example, genetic engineering.

Methodologically, the most common research strategy was that of the case study, which was employed in just over half the studies. Twenty-eight studies reviewed used experimental designs seeking to explore the effects of small-group discussions compared with other approaches.

The most popular techniques for gathering data were observation, video- and audio-tapes of discussions and interviews. Questionnaires and test results also were used.

It is not surprising that more than half the studies employed case studies, because a characteristic of work in the area is a desire to gather detailed information about the nature of discussions. One outcome of the case-study approach and the very labour-intensive nature of much of the data collection and analysis was that sample sizes tended to be small – very often one class or one or two groups of students within a class. Studies involving several

classes or classes in more than one school were comparatively rare.

In contrast to the review of context-based/STS approaches, the studies included in the in-depth review were not limited to experimental studies but focused on the medium- to high-quality studies, of which there were twenty-five. In reaching decisions about quality in a review that encompassed a substantial number of qualitative studies, the Review Group for Science drew on the work of Spencer et al. (2003), who had developed a framework for assessing qualitative research evidence in response to some of the criticisms of EPPI-Centre review methods. This area is explored in more detail in the next section of this chapter.

EVIDENCE ON THE NATURE OF SMALL-GROUP DISCUSSIONS

Nineteen studies addressed the nature of small-group discussions, and the evidence reported here is based on the fourteen studies rated as medium quality or better.

The review revealed a number of features of particular interest in relation to the use of small-group discussion work in science. It is clear from the studies that a complex and interacting set of factors is involved in enabling students to engage in dialogues in a way that could help them to draw on evidence to articulate arguments and develop their understanding. Thus a particular characteristic of such studies is very detailed description of student interactions.

Although the studies in the in-depth review shared a number of similar characteristics at the broad level, there are considerable differences at the detailed level. There was considerable variety in the specific research questions, the topics used for the discussion tasks and the use and interpretation of the term 'small-group discussion'. It was apparent that small-group discussions were being used in a variety of ways in science lessons, with many of the studies wrapping up small-group discussions within other activities, often characterised as 'collaborative learning'. This term itself was used in a variety of ways, often loosely, and on occasion, it appeared to include most activities which did not involve teacher exposition. Despite this variety, there is a high degree of consistency in the findings and conclusions. In general, students often struggle to formulate and express coherent arguments during small-group discussions and demonstrate a low level of engagement with tasks.

The review presents very strong evidence of the need for teachers and students to be given explicit teaching in the skills associated with the development of arguments and the characteristics associated with effective group discussions. Indeed, five of the seven highest-quality studies in the review make this recommendation.

The review presents good evidence that groups function better when the stimulus used to promote discussion involves both internal and external conflict, that is, where a diversity of views and/or understanding is represented within a group (internal conflict) and where an external stimulus presents a group with conflicting views (external conflict).

There is good evidence on group structure. Groups functioned better when they were specifically constituted such that differing views were represented. There is also evidence to suggest that assigning managerial roles to students (e.g., reflector, regulator, questioner or explainer), as suggested in collaborative learning theory, is likely to be counter-productive for poorly structured tasks.

Some evidence also was presented which suggests that single-sex groups may function better than mixed-sex groups, although overall development of understanding was not affected by the group gender composition. Group leaders emerged as having a crucial role: Those who were able to adopt an inclusive style, and one which promoted reflection, were the most successful in achieving substantial engagement with the task. An alienating, overly assertive leadership style generated a lot of 'off-task' talk and low levels of engagement.

Finally, little systematic data have been gathered on the effects of small-group discussions on students' attitudes towards science. Methodologically, the review also helped to

provide information on the research strategies adopted to explore aspects of small-group discussion work. A number of similarities emerged in the approaches adopted in the studies. They tended to make use of opportunistic samples, drawing on the researchers' personal contacts. Experimental designs were not used often, although studies often made comparisons between discussion groups in the same class or within a discussion group. Data-collection methods typically involved audio and/or video recordings, with analysis and reporting drawing heavily on extracts from recorded dialogue. Whilst approaches to gathering data were seldom justified in any detail by the authors, sound procedures appeared to have been introduced to check the reliability of the data analysis and present the findings in a way which made them trustworthy. A key difference that emerged concerns the two contrasting approaches to data analysis, with some studies developing grounded theory from the data and others drawing on existing models to structure their analysis.

EVIDENCE ON THE EFFECTS OF SMALL-GROUP DISCUSSIONS ON STUDENTS' UNDERSTANDING OF EVIDENCE

Fourteen studies were included in the in-depth review, of which twelve were medium quality or better. The review suggested that there is reasonable evidence that use of small-group discussions based on a combination of internal conflict (i.e., where a diversity of views and/or understanding are represented within a group) and external conflict (where an external stimulus presents a group with conflicting views) resulted in a significant improvement in students' understanding of evidence. Where there was either internal or external conflict, there was some improvement in students' understanding.

Improvement in students' understanding of evidence correlated with the initial *dissimilarity* of the group members in terms of their understanding of the science content of the discussion task; that is, student groups were constructed such that they contained students with as wide a range of understandings as possible.

Improvements in understanding were independent of gender composition of groups, although single-sex groups functioned more purposefully.

Students' understanding of evidence improved when they were provided with specific guidance on how to engage in small-group discussions and/or construct arguments. There also was some evidence to suggest that the use of small-group discussions (together with specific instruction in argumentation skills) improved students' ability to construct more complex arguments.

The review of the effects of different stimulus materials on understanding of evidence did not suggest that any particular stimulus materials were more effective than others. Rather, it affirmed the findings of the other reviews about the need for guidance on how to engage in the task and that tasks involving internal and external conflict tended to be more successful.

The Possibilities: What Evidence Might Have to Offer to Science Education

Engaging in the process of undertaking systematic reviews has raised a number of issues and questions relating to the review process, to the dissemination of review findings and to the implications of systematic review work for several more general aspects of educational research. This section draws on the experience of the EPPI-Centre Review Group for Science to explore the more conceptual issues and assess the potential of systematic reviews for science education.

General Methodological Issues Associated with Systematic Reviews

There have been a number of challenges associated with the undertaking of systematic reviews in science education. The EPPI-Centre systematic review process is, in the view of the Review Group for Science, relatively non-contentious up to and including the point of developing the systematic map. For the science reviews themselves, many

aspects of the review process were fairly straightforward. There was consensus in the review group over the priority areas for review. There were few problems identifying studies in the areas reviewed, with over two-and-a-half thousand emerging from initial searches and some several dozen meeting the inclusion criteria in each case. (These figures were typical of reviews carried out in other areas of education.) The principal problem encountered in screening the studies was the quality of many of the abstracts: In a substantial number of cases, insufficient information was provided in the abstract to decide if the study met the inclusions criteria, necessitating obtaining and reading the whole paper to make a decision, adding substantially to the time taken to undertake the review. There is certainly a message here for the compiling of good-quality abstracts. The review process was reassuring in that there was a high degree of consensus in the quality-assurance steps incorporated into the review process, whereby more than one member of the review team conducted the same task (e.g., coding the studies or extracting the data). However, it is worth noting here that a much greater consensus was obtained when the quality assurance was undertaken by subject specialist than, for example, by a subject specialist and a non-specialist, although with review experience, with discrepancies arising from the non-specialist's lack of knowledge of the area. This points to the desirability of reviews being undertaken by people with specialist knowledge.

The main problems with the review process have been in working with the coding tools developed by the EPPI-Centre such that they encompass the full range of work in the field and in synthesising the results. Unsurprisingly, given the background against which they were developed, the systematic review tools, coding schemes and processes provided a better 'fit' when applied to quantitative experimental studies than to more qualitative studies. Thus, in the case of the two areas reviewed by the Review Group for Science, the EPPI-Centre methods were easier to apply to the review of the effects of context-based approaches than to the review of the use of small-group discussion work. Clearly, at one level, this issue could be addressed without too much difficulty by making revisions to the tools used in the review. However, problems arise if the underlying philosophy of the review methods is one that places a premium on experimental studies and their findings. Certainly, proponents of systematic reviews have argued that this should be the case (see, e.g., Oakley, 2000; Torgerson & Torgerson, 2001).

There is a sense in which the debate over the sorts of research that should be included in systematic reviews and the value of RCTs and experimental research designs are reminiscent of that of the 1970s on the relative merits of experimental approaches and 'illuminative evaluation' (Parlett and Hamilton, 1972). Now, as then, there are those who believe that experimental research is the principal means by which the 'what works' question can be answered, whilst there are others who feel that it is largely inappropriate for research in educational contexts, much of which is carried out in an environment where the researcher cannot control all possible variables and therefore needs to draw on a wider range of strategies to offer insights and explanations. Now, as then, it is unlikely that a consensus will be reached, and the debate will continue. It is worth noting that current systematic review work appears to be establishing that there are comparatively few examples of RCTs in educational research. Certainly there appear to be very few in science education, where, arguably, the nature of the focus might lead one to expect more in the way of experimental approaches to research.

Paradoxically, an emphasis on experimental studies as higher-quality studies has the effect of distorting systematic reviews in a number of ways. First, it can steer those conducting review to formulate review research questions which are likely to yield experimental studies in the searches. Second, and as a consequence of embarking on reviews more likely to yield experimental studies, the reviews run the risk of generating findings that are not context-sensitive, where such sensitivity might be important.

For example, the United States has a much stronger tradition of experimental research than many other countries. There are examples of reviews conducted in the United Kingdom which have reviewed only studies undertaken in the United States. Whilst is some cases this might not be important, there are contextual features of the educational system in the United States that are very different from the United Kingdom. 'What works' in one country may not work or may work in different ways in another. Third, systematic reviews become unsystematic if only certain types of studies are seen as high quality when findings are synthesised.

Implications of the Review of Context-Based/STS Approaches in Science Teaching

The evidence presented in the review of context-based/STS approaches supports the notion that the use of contexts as a starting point in science teaching is an effective way to improve attitudes towards school science whilst, at the same time, not resulting in any drawbacks in the development of understanding of science ideas. However, the process of conducting the review suggests that there is a range of contextual information within which this very general finding needs to be interpreted.

The review focused on evaluations with experimental designs, and the review group was interested to see how many RCTs emerged, given that these have been described as the 'gold standard' (Torgerson & Torgerson, 2001) of research design and provide the strongest evidence of 'what works' (Oakley, 2000). It is interesting that only one of the studies in the review was an RCT, and this poses the question, Why was this approach so seldom employed?

Certainly, there are practical constraints which may make RCTs less feasible in educational contexts, particularly in relation to the evaluation of large-scale curriculum interventions. Decisions on participation in such interventions rarely can be made by researchers, making it difficult to allocate students or classes randomly to groups that will or will not receive an intervention. Most often, the research design has to be built around existing class sets in schools. In the review of context-based approaches, the sampling often was opportunistic in that schools and classes using a new intervention were identified, and then other schools using more conventional course were identified to create a comparison group of roughly similar size. Practical constraints also frequently made it necessary to gather data from intact classes, and this raises issues to do with the construction of matched samples for control and experimental groups.

The constraints just outlined point towards the use of 'design experiments' as being potentially fruitful (see, e.g., Brown, 1992; Collins, 1993; Cobb et al., 2003). A 'design experiment' in educational contexts involves evaluating the effects of an intervention in a limited number of settings. For example, this might involve selecting teachers who teach roughly comparable groups but who have different teaching styles and exploring the effects of the intervention on each group of students. The design experiment then yields information on the circumstances in which the intervention is likely to be most successful. Design experiments have the advantage of being able to encompass the complexity of educational settings whilst enabling the aims of interventions to be tested systematically.

A more fundamental point about the use of RCTs concerns the 'What works?' question, which is not as simple as it first appears in the context of the evaluation of an educational intervention. Before it is possible to decide 'what works', it is necessary to decide what 'working' means – and 'working', quite legitimately, may mean different things. This can be illustrated with reference to the study mentioned earlier in which students following the context-based course performed significantly less well on standard test items of chemical knowledge and understanding than students following more conventional courses. However, students in both groups achieved similar grades in their final examinations, where standards are rigorously

monitored to ensure comparability, but students are examined through styles of questions that most closely resemble the teaching and learning approaches, that is, context-based questions for student following a context-based course and conventional questions for students following conventional courses. Thus, if 'what works' means getting similar marks on traditional-style questions, the context-based course clearly does not 'work'. However, if it means getting similar grades on external examinations judged to be of the same standard, then it does 'work'. It seems perfectly reasonable to suggest that if the aims of an intervention are different, the way that it is evaluated will need to be different such that judgements are reached as declared aims and not by comparisons with another approach. In this specific case, most context-based/STS interventions involve a shift in the intended outcomes for science teaching, and the old and the new therefore cannot be compared directly, making the 'What works?' question more problematic to answer.

Three weaknesses were most apparent in the research on context-based approaches. These were lack of standardisation of instruments, the matter of who collected the data and for what purpose and the nature of the resources.

Each of the studies reviewed employed different instruments to gather data on attitudes and/or understanding. This variety meant that it was not feasible to make direct comparisons between studies or to undertake any meta-analysis of the data. This raises the question of how feasible it might be to make use of standardised instruments in the evaluation of context-based approaches when such approaches are developed and used in a number of countries. There would appear to be scope for the development of a standardised instrument to measure attitudes, although research in the area has been characterised for several decades by a tendency for new instruments to be developed for specific studies or existing instruments to be adapted for use. However, there would be considerable merit in trying to put together a small bank of well-validated instruments on which researchers might draw when wanting to assess attitudes towards science or have a common 'core' of items to be included. Cross-national tests of understanding are more problematic. Those developing items for use in international assessments of understanding in science (and other areas), such as The International Mathematics and Science Survey (TIMSS) and the Programme for International Student Assessment (PISA), have encountered a number of challenges. Countries differ in their educational frameworks in relation to when students start school, to the number of years of compulsory schooling, to the ages that students sit national tests and examinations and in the curriculum students have experienced by these points. All these factors mitigate against the validity of using some form of cross-national measure of scientific understanding, although the problems would appear to diminish as pupils reach their final years of schooling and are more likely to have covered the full range of areas common to school science curricula.

The matter of who collects the data and for what purposes also raises issues to do with the quality of the research. It was very noticeable that this information was difficult to identify in the studies and, in almost all cases, had to be drawn by inference. Two particular patterns emerged. The most common situation was for study authors to have been involved in the development of the intervention as well as evaluation its effects. Although it was not clarified, such studies appeared to be undertaken for personal interest rather than to satisfy any funders/sponsors. In other cases, the study authors were users of the intervention and collected their data for personal interest as part of their studies for a higher degree. However, there was an absence of independent, external evaluation. The involvement of developers and users in the evaluation does raise ethical issues about introducing possible bias into the evaluation findings because it could be argued that developers have a vested interest in demonstrating that their intervention has been successful. However,

this appears to be less of an issue than might be the case because detailed examination of the studies during the review process suggested that the appropriate steps were taken to minimise such bias.

Turning to the nature of the resources, the information in the studies included in this review focused on the evaluation data, and very few, if any, examples of the resources were included. It is clear from the studies that the terms 'context-based approaches' and 'STS approaches' can be interpreted quite broadly. This suggests that some caution is needed in interpreting the findings of the review because it seems difficult to imagine that all contexts have the same effects on all students. However, this caveat can be set against a background of the consistency of the evidence yielded by the studies taken as a whole.

Implications of the Review of Small-Group Discussions in Science Teaching

The review suggests that small-group discussion work can provide an appropriate vehicle for assisting students in the development of ideas about using evidence and constructing well-supported arguments. As with the review of context-based approaches, this general finding needs to be interpreted within a range of contextual information.

Two particularly striking features emerge from the work undertaken for the review in relation to the nature of the research and the approaches to analysis. First, it is very apparent that there is considerable variation in the nature of research into small-group discussion work, particularly in relation to its focus, the clarity with which any variables being investigated are specified and the techniques used to analyse data. Second, two very contrasting approaches to data analysis emerged, with some studies developing grounded theory from the data and others drawing on existing models to structure their analysis. This finding suggests that research into small-group discussions in science teaching would benefit from a consideration of discourse analysis techniques developed in other subject areas, such as English, to establish what they might have to offer work in science.

The review also revealed considerable uncertainty on the part of teachers as to what they are required to do to implement good practice. Given that current policy strongly advocates the use of small-group discussion work, both these factors point to a pressing need for a medium- to large-scale research study which focuses on the use and effects of a limited number of carefully structured small-group discussion tasks aimed at developing various aspects of students' understanding of evidence and that such a study should be linked to a coherent analysis framework. This work then could very usefully inform the nature of guidance offered to teachers and students on the development of the skills necessary to make small-group discussions work effectively. This, in turn, points to the desirability of professional development training for teachers.

Methodologically, the review of small-group discussions also demonstrated some of the limitations of the EPPI-Centre review process when extracting data from studies. As noted earlier, the review contained a high proportion of descriptive studies. Such studies were felt by the Review Group for Science to be a valuable source of information in an area of work in its infancy in science education. The review group therefore decided to enhance the EPPI-Centre data-extraction process by drawing on the guidance for assessing research evidence in qualitative research studies developed by Spencer et al. (2003). This enhanced data extraction addressed such matters as details of data-collection methods (including the rationale for their development), measures taken to increase the trustworthiness of the data, information provided about descriptive analytical categories generated and used, diversity in the data, trustworthiness of the analysis process, information provided about the generation of criteria for effectiveness or impact and the relatability of findings. The review group felt that the addition of these questions greatly enhanced the data-extraction process and the quality of the information yielded.

Full details of this process may be found in Bennett et al. (2004b).

Summary and Conclusions

In terms of the substantive findings of the reviews, the Review Group for Science has concluded that the reviews it conducted have provided insights into effects rather than concrete evidence of 'what works'. Reviews are based on a limited number of studies, many of which are small scale and unrelated to other studies. Taken together, these three factors mean that the evidence base is unlikely to be extensive, and any recommendations made would need to be seen as tentative hypotheses to be tested through the gathering of empirical data. Thus review findings cannot currently form a particularly secure basis on which to make recommendations for evidence-informed policy and practice. Rather, the reviews clear the ground by providing a picture of the current state of work and set an agenda for further research which could inform decisions about policy and practice.

There are a number of benefits associated with the review process and products. The systematic map of work in the field, based on all the studies that meet the inclusion criteria, represents a very valuable resource in terms of both systematically identifying and characterising research undertaken in an area and pointing to under-researched areas. There are also considerable benefits to be had from following systematic review methods in any review of research, although the full review process is very resource intensive. The Review Group for Science does not believe, as proposed by Torgerson (2003), that every piece of primary research should be preceded by a systematic review. The inherent resource implications would appear to make this largely impractical in many situations. However, streamlined reviews are possible, and the experience of members of the Review Group for Science has demonstrated that the ability to offer such reviews is attractive to research funders.

More widely, there is a positive outcome from the systematic review/RCT debate, which is to encourage the educational research community to look more closely at the possibility and desirability of undertaking RCTs in educational research. What seems to be important here is not to see RCTs as some 'gold standard' of research design but to ask the question, *When* is such a technique appropriate? It will be interesting to see in the next few years the extent of the impact of RCTs on the design of research studies in education.

The Review Group for Science has considered the 'objectivity' of the review process, particularly in relation to comments about the relative merits of systematic and narrative reviews. The experience of the review group suggests that reviews are less objective than they purport to be. Value judgements, many of which are rooted in professional experience, are inherent in several aspects of the review process, including specification of the inclusion criteria and differences in interpretation of material in abstracts and studies, not all of which are simply related to depth and breadth of knowledge of the field. Certainly the final synthesis of quality is a far from mechanical process, drawing very heavily on knowledge and expertise in the field. Thus the *process* is transparent and replicable, but the *products*, including the systematic map and the in-depth review, depend on the values held and judgements made by those involved in the review process.

The Review Group for Science also has given some thought into what the process of engaging in the reviews has suggested about the quality of educational research. Based on their experience, review group members would support some of the criticisms made by Hargreaves (1996). There is a sense of much research not drawing on previous work or being cumulative in nature. The review of small-group discussion work in particular suggested that disparate approaches were being adopted by studies which were ostensibly addressing similar questions, resulting in a fragmented picture of the overall findings in an area rather than

a cohesive evidence base from which suggestions and recommendations for policy and practice could be made. Whilst these messages may not reflect the situation in all areas of research in science education, they would seem important for the science education community to consider.

In conclusion, systematic review methods appear to have much to offer educational research and research in science education. They may not yet have reached a point where they can answer existing questions about learning and 'what works' except in certain limited areas. However, they have the potential to make a very valuable contribution to research through providing a firmer foundation for decisions about future empirical research; through improving the comprehensiveness, clarity and rigour of research studies; through contributing to the establishment of a culture of evidence-enriched practice; through stimulating informed debate about the nature, purpose and quality of educational research; and ultimately through contributing to the accumulation of reliable evidence on educational practice.

References

Bennett, J., Hogarth, S., Lubben, F., Campbell, B. & Robinson, A. (2010). Talking science: The research evidence on the use of small group discussions in science teaching. *International Journal of Science Education* 32(1), 69–95.

Bennett, J., Lubben, F. & Hogarth, S. (2007). Bringing science to life: A synthesis of the research evidence on the effects of context-based and STS approaches to science teaching. *Science Education* 91(3), 347–70.

Bennett, J., Lubben, F., Hogarth, S. & Campbell, B. (2005). Systematic reviews of research in science education: Rigour or rigidity? *International Journal of Science Education* 27(4), 387–406.

Bennett, J., Lubben, F., Hogarth, S., Campbell, B. & Robinson. A. (2004a). A systematic review of the nature of small-group discussions in science teaching aimed at improving students' understanding of evidence. In *Research evidence in education library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Bennett, J., Lubben, F., Hogarth, S. & Campbell, B. (2004b). A systematic review of the use of small-group discussions in science teaching with students aged 11–18, and their effects on students' understanding in science or attitude to science. In *Research evidence in education library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Bennett, J., Lubben, F. & Hogarth, S. (2003) A systematic review of the effects of context-based and Science Technology-Society (STS) approaches to the teaching of secondary science. In *Research evidence in education library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Brown, A. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences* 2(2), 141–78.

Cobb, P., Confrey, J., diSessa, A., Leherer, R. & Scauble, L. (2003). Design experiments in educational research. *Educational Researcher* 32(1), 9–13.

Cochrane, A. (1972). *Effectiveness and efficiency: Random reflections on the health services*. London: Nuffield Provincial Hospitals Trust.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Collins, A. (1993). Toward a design science of education. In E. Scanlon and T. O'Shea (eds.), *New directions in educational technology*. New York: Springer-Verlag.

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews*, 3rd ed. Thousand Oaks, CA: Sage.

Davies, P. (2000). The relevance of systematic reviews to educational policy and practice. *Oxford Review of Education* 26, 365–78.

Davies, H., Nutley, S. & Smith, P. (eds.) (2000). *What works? Evidence-based policy and practice in public services*. Bristol, UK: Policy Press.

Driver, R., & Bell, B. (1985). Students' thinking and the learning of science: A constructivist view. *School Science Review* 67(240), 443–56.

Evans, J., & Benefield, P. (2001). Systematic reviews of educational research: Does the medical model fit? *British Educational Research Journal* 27(5), 527–41.

Hammersley, M. (2001). On 'systematic' reviews of research literature: A 'narrative' response to Evans and Benefield. *British Educational Research Journal* 27(5), 543–54.

Hargreaves, D. (1996). Teaching as a research-based profession: Possibilities and prospects. Teacher Training Agency Annual Lecture. Teacher Training Agency (TTA), London.

Hillage, L., Pearson, R., Anderson, A. & Tamkin, P. (1998). *Excellence in research on schools*. Brighton, UK: Institute for Employment Studies.

Hogarth, S., Bennett, J., Lubben, F. & Robinson, A. (2006). The effect of ICT teaching activities in science lessons on students' understanding of science ideas. In *Research evidence in education library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Hogarth, S., Bennett, J., Campbell, B., Lubben, F. & Robinson, A. (2004). A systematic review of the use of small-group discussions in science teaching with students aged 11–18, and the effect of different stimuli (print materials, practical work, ICT, video/film) on students' understanding of evidence. In *Research evidence in education library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Levinson, R., & Turner, S. (2001). *Valuable lessons: Engaging with the social context of science in schools*. London: The Wellcome Trust.

Lubben, F., Bennett, J., Hogarth, S. & Robinson, A. (2004). A systematic review of the effects of context-based and Science-Technology-Society (STS) approaches in the teaching of secondary science on boys and girls, and on lower ability pupils. In *Research evidence in education library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Millar, R., & Osborne, J. (eds.) (1998). *Beyond 2000: Science education for the future*. London: King's College/The Nuffield Foundation.

Norris, N. (1990). *Understanding educational evaluation*. London: Kogan Page.

Oakley, A. (2000). *Experiments in knowing*. Cambridge, UK: Polity Press.

Oakley, A. (2002). Social science and evidence-based everything: The case of education. *Educational Review* 54(3), 21–33.

OECD (2002). Educational research and development in England: examiners' report. Organisation for Economic Co-operation and Development. Retrievable at www.oecd.org/dataoecd/17/56.

Osborne, J., Duschl, R. & Fairbrother, R. (2002). *Breaking the mould? Teaching science for public understanding*. London: Nuffield Foundation.

Parlett, M., & Hamilton, D. (1972). Evaluation as illumination: A new approach to the study of innovative programmes (Occasional Paper No. 9). Centre for Research in the Educational Sciences, University of Edinburgh, Edinburgh, Scotland.

Petrosino, A., Boruch, R., Rounding, C., McDonald, S. & Chalmers, I. (2000). The Campbell Collaboration: Social, Psychological, Educational and Criminal Trials Register (C2-SPECTR). *Evaluation and Research in Education* 14(3), 206–19.

Shavelson, R., & Towne, L. (eds.) (2002). *Scientific enquiry in education*. Washington: National Academy Press.

Slavin, R. (2002). Evidence-based educational policies: Transforming educational practice and research. *Educational Researcher* 31(7), 15–21.

Spencer, L., Ritchie, J., Lewis, J. & Dillon, L. (2003). *Quality in qualitative evaluation: A framework for assessing research evidence*. London: The Strategy Unit.

Tooley, J., & Darbey, D. (1998). *Educational research: A critique. A survey of published educational research*. London: Office for Standards in Education (Ofsted).

Torgerson, C., & Torgerson, D. (2001). The need for randomised controlled trials in educational research. *British Journal of Educational Studies* 49(3), 316–28.

Torgerson, C. (2003). *Systematic reviews*. London: Continuum.

Vulliamy, G. (2004). The impact of globalisation on qualitative research in comparative and international education. *Compare* 34(3), 261–84.

What Works Clearinghouse (2002). A trusted source of evidence of what works in education. Retrievable at http://ies.ed.gov/ncee/wwc/.