Bringing Science to Life: A Synthesis of the Research Evidence on the Effects of Context-Based and STS Approaches to Science Teaching

JUDITH BENNETT, FRED LUBBEN, SYLVIA HOGARTH
Department of Educational Studies, University of York, York, YO10 5DD, UK

Received 23 May 2006; accepted 23 August 2006

DOI 10.1002/sce.20186
Published online 18 October 2006 in Wiley InterScience (www.interscience.wiley.com).

ABSTRACT: Context-based and science–technology–society (STS) approaches to teaching science in high school have become widely used over the past two decades. They aspire to foster more positive attitudes to science while, at the same time, provide a sound basis of scientific understanding for further study. This paper reviews the detailed research evidence from 17 experimental studies undertaken in eight different countries on the effects of context-based and STS approaches, drawing on the findings of two systematic reviews of the research literature. The review findings indicate that context-based/STS approaches result in improvement in attitudes to science and that the understanding of scientific ideas developed is comparable to that of conventional approaches. The approaches also result in more positive attitudes to science in both girls and boys and reduce the gender differences in attitudes. The paper also considers issues emerging from work in the area in relation to study design and the constraints which may militate against the use of experimental research designs when gathering evidence of impact of interventions. A fundamental constraint is the extent to which it is possible to make comparisons between existing methods and interventions when the aims are overlapping but also differ in significant ways. © 2006 Wiley Periodicals, Inc. *Sci Ed* **91:**347–370, 2007

INTRODUCTION

One of the most discernible trends of the last two decades in science curriculum development across a number of countries has been to use contexts and applications of science as a means of developing scientific understanding. Teaching in this way is often described as adopting a *context-based* or *STS* (*science–technology–society*) approach. The trend toward the use of context-based/STS approaches is apparent across the whole age spectrum from

Correspondence to: Judith Bennett; e-mail: jmb20@york.ac.uk
This paper was edited by former Editor Nancy W. Brickhouse.

primary through to university level, but is most noticeable in materials developed for use in the secondary age range.

Many people involved in curriculum development and teaching believe that there are considerable benefits associated with context-based/STS approaches. However, it raises a number of interesting questions: Does teaching science through the use of everyday contexts help school students understand science any better? Does teaching science in context improve school students' attitudes to science? Are there differences in the effects on girls and boys, or on students of different ability?

This paper examines in detail the research evidence on the effects of context-based and STS approaches to the teaching of science. In particular, it looks at the effects on students' understanding of science and on their attitudes to science. The evidence presented has been gathered using the systematic review methods based on those developed as part of the Evidence, Policy and Practice Initiative (EPPI), a Government-sponsored project in the United Kingdom whose aim is to synthesize and disseminate research findings in key areas of education. The paper also summarizes the debate over the utility of systematic review methods in education.

The following definitions have been adopted for the purposes of this paper:

> *Context-based approaches* are approaches adopted in science teaching where contexts and applications of science are used as the *starting point* for the development of scientific ideas. This contrasts with more traditional approaches that cover scientific ideas first, before looking at applications.
>
> *STS* (*science–technology–society*) is a term that is used very broadly, and this paper uses the definition provided by Aikenhead (1994)
>
>> STS approaches [are] those that emphasise links between science, technology and society by means of emphasising one or more of the following: a technological artefact, process or expertise; the interactions between technology and society; a societal issue related to science or technology; social science content that sheds light on a societal issue related to science and technology; a philosophical, historical, or social issue within the scientific or technological community. (p. 52–53)

In presenting these definitions, it is accepted that that there are considerable areas of overlap. The term "context based" appears to be more common in Europe, while "STS" is preferred in North America.

Both context-based and STS approaches make frequent reference to the term *scientific literacy*, an area which is receiving increasing attention in the school science curriculum. In the context of this paper, scientific literacy has been taken to encompass the knowledge, understanding, and skills young people need to develop in order to think and act appropriately on scientific matters that may affect their lives and the lives of other members of the local, national, and global communities of which they are a part.

THE AIMS OF CONTEXT-BASED/STS COURSES

A variety of aims has underpinned the development of context-based/STS materials. Arguably, the most significant of the aspirations lies in the area of students' *affective* responses to science—how they *feel* about their experiences of science. The hope is that the contexts used to develop scientific ideas will motivate students and make them feel more positive about science by helping them see the importance of what they are studying. Interestingly, the earliest examples of courses adopting such approaches were those developed for two quite different groups for whom science had little appeal: nonscience students in tertiary education in the United States and less academic students in secondary schools in the United Kingdom. There is also the hope that, for some students at least, this interest will be translated into a desire to study science subjects beyond the period when they are compulsory. For it is certainly the case that there has been—and continues to be—widespread concern in a number of countries over the comparatively low numbers of young people electing to study science.

A second aim is to do with the learning of scientific ideas: if students are more interested and motivated by the experiences they are having in their lessons, this increased engagement might result in improved learning. The development of understanding of scientific ideas poses a particular challenge for context-based/STS approaches because of the implications for the way that scientific knowledge is introduced. If ideas are introduced as they arise in particular contexts—in other words, on a "need to know" basis—then it is unlikely that any one concept area will be introduced and developed in full in one particular context, as might be the case in more conventional courses. At worst, this might mean students following context-based/STS courses develop a poorer understanding of science.

A systematic review of the research evidence in these areas is therefore particularly pertinent.

SYSTEMATIC REVIEWS IN THE FIELD OF EDUCATION

Systematic reviews of research studies are a comparatively recent development in education, though they are well established in medical research. They have emerged from the international debate over the nature and purpose of educational research, and how it contributes to maximizing the effectiveness of educational provision (e.g., Hargreaves, 1996, and Hillage, Pearson, Anderson, & Tamkin, 1998, in the United Kingdom; Kaestle, 1993, and Shavelson & Towne, 2002, in the United States).

One significant outcome of the debate was that initiatives were made in the early 2000s to introduce elements of the medical model into educational research. The first of these initiatives was the setting up of the Campbell Collaboration in Philadelphia in the United States to review evidence from randomized controlled trials (RCTs) in education, criminology, and other social sciences (see, for example, Petrosino, Boruch, Rounding, McDonald, & Chalmers, 2000). The second was the establishing of the Evidence for Policy and Practice Initiative Centre (EPPI-Centre) in the United Kingdom to focus on systematic reviews of research evidence in key areas of education. The work reported here forms part of this systematic review program.

Proponents of systematic reviews of research emphasize their usage as a source for informing policy decisions and changes in practice. Initially, these reviews aimed to answer questions on "what works?" Several characteristics of the systematic review process are seen as strengths of the findings emerging from systematic reviews (Torgerson, 2003), i.e. the explicit criteria for selecting studies for the review; the exhaustive coverage of studies published on the chosen review topic; the generic criteria for judging quality of the research process; and the quality assurance through the involvement of at least two researchers in decisions of selection, classification, and quality. Consequently, the systematic review process is seen as reliable, valid, and unbiased, and the synthesis of the findings is considered fair as it is weighted according to the strength of the evidence from each study in the review. Also, a review can be replicated and updated easily. The quality assessment has tended to favor experimental studies as providing more rigor (Borman, 2002), in particular the RCTs which have been advocated as the "gold standard" in research literature (Torgerson & Torgerson, 2001) and in guidelines for practitioners (U.S. Department of Education, 2003).

The association of systematic reviews with positivist approaches to educational research has resulted in considerable criticism being leveled against systematic reviews as a research tool, much of which resonates with the critique of quantitative research methods in the 1970s and 1980s. Hammersley (2001) suggests that the systematic review methods and the high status of RCTs in the quality assessment give privilege to a positivist perspective of educational research. Part of his objection is based on the focus on educational outcomes, i.e. effectiveness and impact, when review questions are limited to the type "what works?" A positivist perspective is also considered inappropriate because of the complex nature of settings of educational research which rarely show a singular cause–effect linkage even after consciously controlling for other factors, as unpredicted factors also impact on the targeted outcome. These "complexity" objections come from researchers using qualitative methods to explore educational processes (for instance, Elliott, 2001) and also from researchers using quantitative methods based on the complexity theory (Byrne, 2005). Others see systematic reviews as an outcome of globalization and the inherent threat of the loss of diversity of educational research (Vulliamy, 2004).

A second body of critique on systematic reviews of research literature focuses on the relationship between research findings and educational practice. Burkhardt and Schoenfield (2003) distinguish three research traditions within education. The *humanities* approach aims at the generation of ideas which are assessed for their plausibility, internal consistency, and the fit with current theories. Critical commentary is the product. The *science* approach aims at "improved knowledge of 'how the world works' through the analysis of phenomena and the building of models to explain them" (p. 5) with empirical tests providing evidence for assertions. Lastly, the *engineering* approach aims at solving practical problems using knowledge of how the world works to "make the world work better," through design experiments— designing and testing of educational products and processes. The result includes products and processes with evaluated evidence for their effectiveness for specific users. Most proponents of systematic reviews will support a *science* approach to educational research—the synthesized findings will provide evidence for incorporating or informing practice. However, systematic reviews also display many of the key features of the *engineering* approach described above: rather than generating a finished research product, they inform "design experiments" (Brown, 1992) by pointing to ways in which hypotheses may be tested in specific contexts for specific users.

For the authors, the review reported here served two purposes. First, it enabled the synthesis of research evidence in an area of considerable interest to science educators. Second, it provided a valuable opportunity to explore the potential utility of systematic reviews in education and, specifically, gather data to indicate the extent to which experimental methods were used to evaluate the effects of intervention strategies.

SYSTEMATIC REVIEW METHODS

This section provides a brief overview of systematic review methods. A more detailed account and critique of systematic review methods may be found in Bennett, Lubben, Hogarth, and Campbell (2005).

The systemic review process, as developed by the EPPI-Centre, involves several stages: identification of review topic area; identification of *review research question*; development of *inclusion and exclusion criteria* for studies in the review (relating to, for example, aspects such as the age of students, the nature of the research design, and the reported outcomes); undertaking of *systematic searches* of electronic databases and other sources for potentially relevant research studies; coding or *keywording* studies against prespecified and agreed characteristics; production of an overview or *systematic map* of studies in the review area that groups the studies according to their chief characteristics; undertaking an *in-depth review* of studies to look in detail at their design and findings and to evaluate the quality of the work reported.

The in-depth review involves a process called data extraction, where information from the studies is extracted in a systematic way. Information extracted from the studies includes study aims and rational; study research questions; study design methods, including selection of groups, sampling, and consent of subjects; data collection methods; data analysis methods; reliability and validity of methods of data collection and analysis; results and conclusions; quality of reporting; quality of the study in relation to methods and data. This information is then used to make judgments about the quality (high, medium, or low) of the weight of evidence presented in the study in relation to the review research question.

Peer-refereed reports of reviews are available in electronic form through the Research Evidence in Education Library (REEL), an open-access resource available through the EPPI-Centre Web site (http://eppi.ioe.ac.uk).

Stakeholder participation forms a key element of reviews in order to ensure their topicality and relevance. Thus their work is overseen by a *review group*, whose membership includes policy makers, teachers, inspectors, academic researchers, teacher trainers, and those involved in curriculum innovation, including textbook authors.

TEACHING SCIENCE IN CONTEXT: THE REVIEW RESEARCH QUESTION

The review research question developed for the work reported here is: *What evidence is there that teaching approaches that emphasize placing science in context and promote links between science, technology, and society (STS) improve the understanding of science and the attitudes to science of 11–18-year-old students?*

Studies included in the review met the following criteria:

- Their principal focus is on the effects of context-based/STS approaches on students understanding of science or attitudes to science.
- They have been undertaken with students aged between 11 and 18.
- They report evaluations of context-based/STS materials.
- They have been published in English and in the period 1980–2003.
- They are reported in journal articles, books or collected works, conference proceedings, dissertations, or reports.

Student age was restricted to 11–18 because the majority of context-based/STS curriculum development projects have been aimed at this age range. The start date for the period of publication was dictated by the fact that the earliest examples of context-based/STS materials date from the beginning of the 1980s. The initial search was cast widely (and therefore included conference papers and reports) in order to increase the validity of the review of the field, with stringent quality criteria being applied at a later stage of the review, i.e. in the in-depth review. The searches of ERIC, BEI, SSCI, and PsychINFO databases yielded some 2500 studies, of which 61 met the inclusion criteria for the review.

These studies were then coded against particular characteristics (*keyworded*) to produce an overview (*the systematic map*) of evaluation studies on context-based/STS interventions. In producing the map, the following characteristics of studies were scrutinized: the country of study; the age/level of the students; the type of study; the discipline of the study; the nature of the intervention; the outcome measures; the outcomes.

The next section of the paper presents an overview (systematic map) of the 61 evaluation studies that met the inclusion criteria for the review. Following the discussion of the systematic map, the review progressively focuses down to, first, the 24 studies that adopted an experimental design, and second, the 17 experimental studies that were judged to be of good quality (see later Discussion). This enabled the in-depth review to focus on good quality studies yielding evidence of impact on understanding and attitude.

AN OVERVIEW (SYSTEMATIC MAP) OF EVALUATION STUDIES ON THE EFFECTS OF CONTEXT-BASED/STS APPROACHES

Just over 80% of the studies in the systematic map were carried out in the United States, the United Kingdom, the Netherlands, and Canada, with others being carried out in Estonia, Germany, Ireland, Israel, Scotland, Swaziland, and Taiwan.

There were a number of productive research groups in the area. A minimum of three studies from the same source give some indication of clusters of activity and, in the first four listed below, point to coordinated research programs. In particular, there were clusters of reports of studies undertaken by researchers at the University of Utrecht in the Netherlands on the Dutch Physics Curriculum Project (PLON), researchers at the University of York Science Education Group (UYSEG) on the Salters curriculum development projects and context-based materials developed for use in Southern Africa, Yager and coworkers in Iowa in the United States on a variety of aspects of the use of STS materials, and Zoller and coworkers in Israel and British Columbia in Canada on STS courses for upper high school students.

Forty-one studies were undertaken with students in the 11–16 age range, and the remainder with students in the 17–18 age range. The emphasis on students in the 11–16 age range is likely to reflect the perception of this age group being very critical in terms of interest in science declining.

All the 61 studies were evaluations, with 24 employing experimental research designs, i.e. using some form of control group. The remainder explored effects only on students experiencing the context-based/STS materials.

Just over half the studies (35) focused on initiatives characterized as science. Where there was a single-subject focus, 13 related to chemistry, 10 to physics, and 3 to biology. The focus on chemistry and physics in the individual science disciplines is a reflection of the motives for developing context-based materials in the first instance: attitudinal surveys spanning three decades have shown that attitudes to the physical sciences are much more negative than the biological sciences (see, for example, Gardner, 1975; Kelly, 1986; Osborne, Simon, & Collins, 2003).

Interventions were characterized as *context-based* or *STS* on the basis of the terminology used by the authors of the studies. They were further characterized as *full courses* (taken to be program of 1 year or more in duration) or *enrichment materials* (interventions of a shorter duration, typically 2–4 weeks). 39 out of 61 (64%) of the studies evaluated interventions characterized as "context-based" with two thirds of these dealing with full context-based courses. The remaining studies report on evaluations of STS approaches. In contrast with context-based interventions, the majority of evaluations of STS approaches focused on STS enrichment activities within traditional approaches. Of the 61 studies in the map, 24 employed experimental methods in the evaluation, with 18 of these studies focusing on full context-based or STS courses.

Table 1 summarizes the outcome measures in relation to the nature of the evaluation.

The most common outcome measures employed in studies were test results (27 studies), open questionnaires (27 studies), agree/disagree scales (21 studies), and interviews

TABLE 1
Outcome Measures

	Experimental Studies ($n = 24$)	Nonexperimental Studies ($n = 37$)
Test results	15	12
External exam results	0	3
Written reports/open questionnaires	11	16
Agree/disagree scores	10	11
Self reports (e.g., interviews)	4	16
Group discussions	1	1
Presentations	0	1
Observed behavior	3	7
Other	4	1
Totals	48	68

Total is more than 61 as several studies employed more than one outcome measure.

(20 studies). Unsurprisingly, test results were the most commonly used measure in experimental studies, as they were used in almost two thirds of the cases. Questionnaires and interviews featured more prominently in nonexperimental studies, a feature likely to reflect the central purpose of the styles of evaluation.

Forty-four of the studies reported on attitudes and 41 on understanding. Of these, 24 reported on both these aspects. Two other aspects that also emerged as featuring prominently in studies were the effects in relation to gender (17 studies) and low ability (7 studies). It is striking that the effects of gender and low ability are explored almost exclusively for the 11–16 age range, in which science is mostly taken as a compulsory subject.

THE IN-DEPTH REVIEWS

The central aim of the work reported here was to gather research evidence on the effects of context-based/STS approaches on students' understanding of science and attitudes to science. Thus the in-depth review focused on the studies that adopted *experimental research designs*, i.e. comparative data were gathered from a control group following a more conventional science course and an experimental group following a context-based/STS course. Detailed summaries (data extraction) were therefore made of the 24 studies that employed such an approach. While it is recognized that nonexperimental designs have an important role to play through the potential they offer to yield valuable insights into the processes and outcomes of interventions, studies taking the form of appropriately designed experiments are those most likely to yield evidence of effects.

Judgments about the quality of the research design and findings of these 24 studies were then made by at least two researchers on the basis of information in the detailed summaries and, where in doubt, the full papers. Five raters were involved in the quality judgments, and all were trained through a full-day workshop. Interrater agreement on the quality judgments across the 24 studies was 83%, and initial disagreements were easily resolved through discussion. Judging the quality of studies is not a straightforward task, particularly when they report evaluations of complex interventions, such as context-based/STS programs. The criteria that were applied to reach these judgments related to

- the steps taken to establish the reliability and validity of the data collection methods;
- the steps taken to establish the reliability and validity of the data analysis;

- the steps taken to rule out any other sources of error/bias that would lead to alternative explanations for the findings of the study;
- the size of the sample;
- appropriate matching of control and experimental groups;
- appropriate data collection (pre- and postintervention, or postintervention);
- the inclusion of attitude and/or understanding as explicit independent variables;
- the appropriateness of the measures employed to assess understanding and/or attitude;
- the range of outcome measures reported; the representativeness of the situation to that of normal classrooms.

These criteria drew on a range of guidance from published literature on research design. Evidence that informed judgments about the quality of the data collection methods included factors such as the piloting of instruments previously validated in a different context or with a different type of respondent, through the direct use of previously validated instruments down to the use of instruments that had neither been piloted nor validated (Cohen, Manion, & Morrison, 2000). The use of more than one instrument to enable triangulation of findings was also considered to increase the quality of the data, particularly where such instruments appeared to be combining different methods (i.e., quantitative and qualitative), as advocated by, for example, Gorard and Taylor (2004).

For the evaluation of changes in attitudes, any evidence of the development of quantitative items on the basis of previous exploratory interviews with the respondent group was considered an additional indication of quality (Oppenheim, 1992). For the evaluation of changes in conceptual understanding, any measure taken to increase the content validity of instruments, preferably going beyond the simple peer validation for face validity (Black, 1998), was considered a strength in the quality of the data collection process. The appropriateness of the statistical analysis methods and the attention to effect sizes (Rosnow & Rosenthal, 1996) was seen as contributing to the quality of validity of the analysis, just as any strategies adopted to triangulate data from different sources. The involvement of more than one person in the coding and analysis of the same data set increased the reliability of the analysis when accompanied with indicative interrater agreement percentages or Cohen-kappa coefficients (Cohen, 1988).

Evidence for an appropriate match of control and experimental groups included similarity in size, gender balance, and performance on external tests. The extent to which the groups were randomized was also taken as a measure of the quality of the evaluation study. Within this, it was recognized that units of study often are classes rather than individual students, and the criteria proposed by Ukoumunne et al., (1999) of any treatment of less than four classes were applied to the studies in the review.

Studies were judged to be of a higher quality if they involved collecting pre- and postintervention data, rather than only postintervention data, with the appropriate statistical treatment for standardization. Any evaluation of changes in conceptual understanding was seen to be strengthened through the use of a range of outcome measures (e.g., different levels of Bloom's taxonomy of educational objectives, or Piaget's levels of thinking). Similarly, an evaluation of changes in attitudes was seen as more robust if it covered several outcome measures, such as attitudes to science, science learning, science careers, or scientists. Finally, the gathering of data in normal classroom settings and with natural learning groups was seen as a strength as it increases the consequential validity (Messick, 1995) of the findings, i.e. the match between the nature of the data and the intended use of any (evaluative) inferences subsequently drawn.

Of the 24 studies, 17 were judged to be of *medium* (M), *medium high* (MH), or *high* (H) quality. Those studies rated as *medium high* or better were judged to have met most or

TABLE 2
Some Factual Details of the 17 Studies Included in the Detailed Review

	Outcomes Explored and Overall Quality Rating[1]							
Study	Understanding		Attitude		Gender		Ability	
Banks (1997)	✓	MH						
Barber (2000)	✓	MH	✓	M				
Barker and Millar (1996)	✓	H						
Ben-Zvi (1999)			✓	M				
Key (1998)			✓	MH				
Lubben et al. (1997)	✓	M						
Ramsden (1997)	✓	MH						
Rubba et al. (1991)	✓	M						
Smith and Bitner (1993)	✓	M			✓	M		
Smith and Matthews (2000)			✓	M	✓	M		
Tsai (2000)	✓	M						
Wierstra (1984)	✓	MH	✓	MH	✓	M		
Wierstra and Wubbels (1994)	✓	MH	✓	MH				
Winther and Volk (1994)	✓	MH						
Yager and Weld (1999)	✓	H	✓	H	✓	H	✓	H
Zoller et al. (1990)			✓	M	✓	M		
Zoller et al. (1991)			✓	M				
Total = 17	12	H = 2 MH = 6 M = 4	9	H = 1 MH = 3 M = 5	5	H = 1 MH = 0 M = 4	1	H = 1 MH = 0 M = 0

[1] M = Medium, MH = medium high, H = high.

all of the above criteria, while those rated as *medium* met the majority of the criteria. The remaining seven experimental studies were characterized by one or more of the following features: they were small-scale studies with comparatively small sample sizes, there were few reliability and validity checks, and they reported only a narrow range of outcomes.

While EPPI systematic review methods promote the inclusion of all studies, irrespective of quality judgment, the view of the authors of this paper is that it is more productive to focus on those of *medium* or higher quality for the synthesis of study findings. A characteristic of these studies is that they have normally been reported in peer-refereed journals. The remainder of this paper therefore concentrates on the evidence yielded by the 17 studies rated *medium* or better in quality. Full reference details for these studies may be found in the References for the Studies Included in the Review section. Table 2 shows the focus of each of the studies and the overall quality judgments in relation to the evidence they yield about understanding of science, attitudes to science, gender effects, and ability effects.

The Nature of the Interventions

Fifteen of the 17 studies reported evaluations of interventions that took the form of whole courses of a duration of at least 1 year. The remaining two studies gathered data on enrichment modules, i.e. shorter interventions that were not intended to be whole courses.

With the exception of the studies of Ben-Zvi (1999), Rubba, McGuyer, and Wahlund (1991), and Smith and Matthews (2000), where no details are provided, all the interventions received external funding for their development.

The interventions are summarized in Table 3.

TABLE 3
The Nature of the Interventions

Study	Intervention	Country	Notes
Banks (1997), Barber (2000), Barker and Millar (1996), Key (1998)	*Salters Advanced Chemistry*, a 2-year context-based course for students aged 17–18	England	Banks focused on teaching of equilibrium. Key focused specifically on one aspect of the course, the chemical industry visit module
Smith and Bitner (1993), Winter and Volk (1994)	*ChemCom*, a 1-year STS course for high schools students (taught to groups aged between 12 and 17)	USA	The course was offered to nonscience majors aged 15–17
Wierstra (1984), Wierstra and Wubbels (1994)	*PLON (Projekt Leerpakket Ontwikkeling Natuurkunde, Dutch Physics Curriculum Development Project)* project, a 5-year context-based physics course for students aged 12–17	The Netherlands	Wierstra and Wubbels focused on one 4-week module in the course, *Traffic*
Zoller et al. (1990), Zoller et al. (1991)	*STS British Columbia,* a 1-year STS program for students aged 16–17	Canada	
Ben-Zvi (1999)	*Science and Technology For All*, a 1-year STS course for nonscience students	Israel	One six–seven-week module, *Energy and the Human Being,* formed the focus of the study
Ramsden (1997)	*Science: The Salters Approach*, a 2-year context-based science course for students aged 14–16	England and Wales	
Smith and Matthews (2000)	One-year STS course for students aged 14–15 (transition year)	Ireland	
Tsai (2000)	*STS-Taiwan,* a 1-year STS course for students aged 16	Taiwan	Modules on light, electricity, and nuclear energy formed the focus of the study
Yager and Weld (1999)	*Scope, Sequence and Continuity (SS&C),* a 5-year context-based course for students aged 11–16	USA	
Lubben et al. (1997)	*Matsapha* project materials taught to students aged 13–14	Swaziland	A unit covering current electricity formed the focus of the study
Rubba (1991)	STS module taught to students aged 14–15	USA	A module on genetics formed the focus of the study

The Nature of the Evaluations

In contrast to the development of the intervention, comparatively few of the evaluation studies received any funding. Those that gave details of funding for the evaluation were the six studies by Barker and Millar (1996), Lubben, Campbell, and Dlamini (1997), Tsai (2000), Wierstra (1984), Wierstra and Wubbels (1994), and Yager and Weld (1999). This pattern of funding being much more strongly tied to the development of the materials for the intervention, rather than their evaluation, is very typical of context-based/STS programs. Details of the evaluations are given in Table 4.

THE EVIDENCE ON UNDERSTANDING OF SCIENCE

The evidence on understanding of science comes from the findings of 12 studies, 2 of high quality, 6 of medium high quality, and 4 of medium quality (see Table 2 for details).

Data on understanding were gathered either through the use of written answers to diagnostic questions (Banks, 1997; Barber, 2000; Barker & Millar, 1996; Ramsden, 1997) or, more commonly, through the use of examination and test items drawn from existing banks of items (Lubben et al., 1997; Smith & Bitner, 1993; Wierstra, 1984; Wierstra & Wubbels, 1994; Winther & Volk, 1994; Yager & Weld, 1999). Rubba et al. (1991) used a self-developed test of achievement. Only one study (Tsai, 2000) gathered data on understanding through the use of interviews.

Just over half the studies report evidence that indicates context-based/STS approaches develop a level of scientific understanding that is comparable to that of conventional courses (Barber, 2000; Barker & Millar, 1996; Lubben et al., 1997; Ramsden, 1997; Rubba et al., 1991; Smith & Bitner, 1993; Wierstra, 1984; Wierstra & Wubbels, 1994). Four studies indicate that context-based/STS approaches lead to a better understanding of science than in conventional courses (Banks, 1997; Tsai, 2000; Winther & Volk, 1994; Yager & Weld, 1999). In the case of Tsai's (2000) study, students also demonstrated less frequent misunderstandings of ideas. Winther and Volk (1994) comment that the standard test items they used corresponded more closely with the conventional teaching experienced by the control group and therefore felt that their findings were underestimating the gains in understanding developed by the STS approach. Mixed evidence emerged from the Barber (2000) study (see below).

There was some evidence to suggest context-based/STS approaches did not lead to improved understanding. For example, in the study by Lubben et al. (1997), the experimental and control groups were both taught a conventional module, and the experimental group performed significantly better than the control group in the achievement test related to this module. However, when the experimental group received the context-based module, the performance in the achievement test was similar in both control and experimental groups, suggesting that the experimental group performed less well after the context-based intervention.

The findings of two studies (Barber, 2000; Wierstra, 1984) point to a particular issue related to the assessment of understanding when comparing context-based/STS courses with conventional courses that concerns the nature of the items used to provide measures of understanding. The Barber (2000) study reports that students taking the context-based course, *Salters Advanced Chemistry*, got lower scores on the test based on standard assessment items than students taking the conventional course. However, students taking *Salters Advanced Chemistry* did better in their final external examinations. The standard of these final examinations is regulated by an external body, but students taking the context-based course take examinations with context-based questions, rather than more conventional

TABLE 4
Details of the Evaluations

Study	Focus[1]	Sample Details	Summary of Design
Banks (1997)	U	$N = 95$ Control: 17 Experimental: 78 Age: 17–18 Schools: 6	One multipart diagnostic question covering equilibrium Data gathered pre- and postintervention
Barber (2000)	U A	$N = 120$ (A) $N = 35$ (U) Control: 60 (A) Experimental: 60 (A) Control: 20 (U) Experimental: 15 (U) Age: 17–18 Schools: 1	*Understanding*: 14-item test based on test devised by Royal Society of Chemistry; also external examination grades and measures of “value added” *Attitude*: Self-developed questionnaire (the Likert scale and free-response items) plus interviews Data gathered postintervention
Barker and Millar (1996)	U	$N = 140$ Age: 16–18 Control: 70 Experimental: 70 Schools: No details	22 diagnostic questions on elements, compounds, and mixtures; chemical change; conservation of mass in closed and open systems; reacting masses; chemical bonding; thermodynamics; equilibria and rates of reaction Data gathered at three points: start of course, after 7 months, and after 15 months
Ben-Zvi (1999)	U A	$N = 232$ Control: 102 Experimental: 130 Age: 15 Schools: No details	22-item attitude inventory to explore attitudes to science and school science *Note*: Study also used 15 semantic differential scales relating to “image of science” Data gathered postintervention
Key (1998)	A	$N = 1200$ Control: 300 × 3 Experimental: 300 Age: 17–18 Schools: No details	Three questionnaires probing impressions of industry and perceptions of relevance to study of chemistry. *Note*: Study involved comparing *three* conventional courses with context-based course, hence three control groups Data gathered at three points: start-of-course, immediately after visit, and end-of-course

Lubben et al. (1997)	U	*N* = 288 Control: 184 Experimental: 104 Age: 13–14 Schools: 6	Ten questions (standard examination items testing recall, understanding, and application) Data gathered postintervention
Ramsden (1997)	U	*N* = 168 Control: 84 Experimental: 84 Age: 13–14 Schools: 8	Written questionnaire containing eight diagnostic questions on mixtures and compounds, chemical change, conservation of mass Data gathered postintervention
Rubba et al. (1991)	U	*N* = 197 Control: 100 Experimental: 97 Age: 15–16 Schools: 2	Self-developed tests of achievement Data gathered pre- and postintervention and at one other intermediate point in one school and two other points in the other school
Smith and Bitner (1993)	U	*N* = 123 Control: 63 Experimental: 60 Age: 14–16 Schools: 5	Uses Group Assessment of Logical Thinking (GALT) instrument of multichoice items to assess understanding Data also explored for gender effects Data gathered pre- and postintervention
Smith and Matthews (2000)	A G	*N*= 60 Control: 23 Experimental: 37 (questionnaire) Control: 4 Experimental: 8 (interviews) Age: 15–16 Schools: 1	Two questionnaires: (1) 25-item Likert-type questionnaire on perceptions of school science and science; (2) 21-item Likert-type on students' perceptions of science teaching *Interviews*: to gather views on school science, science, and science teachers Data also explored for gender effects Data gathered postintervention

Continued

TABLE 4 ***Continued***

Study	Focus[1]	Sample Details	Summary of Design
Tsai (2000)	U	$N = 101$ Control: 49 Experimental: 52 Age: 15–16 Schools: 1 Data collected from subset of 20 students in each group	Interviews with 20 students selected randomly from each group to establish what they had learned from their instruction. *Note:* Study also used a Likert-type instrument to assess science epistemological beliefs (SEBs) Interview data gathered at three points during the intervention
Wierstra (1984)	U A G	$N = 398$ Control: 144 Experimental: 254 Age: 15–16 Schools: No details	*Understanding*: Physics achievement tests from PLON and traditional physics exams. *Attitude*: 12-item Likert-type questionnaire. Also: 10-item individualized environment questionnaire to assess student perceptions of learning environment Data also explored for gender effects Data gathered postintervention
Wierstra and Wubbels (1994)	U A	$N = 464$ Control: 209 Experimental: 355 Age: 15–16 Schools: No details	*Understanding:* 19-item multiple-choice standard physics tests *Attitude:* 12-item Likert-type questionnaire on responses to school science Also: 10-item individualized environment questionnaire to assess student perceptions of learning environment Data gathered postintervention
Winther and Volk (1994)	U	$N = 93$ Control: 51 Experimental: 42 Age: 15–19 Schools: 1	A standard test of achievement in chemistry was administered pre- and postintervention
Yager and Weld (1999)	U A G Ab	$N = 6590$ Control: 1320 Experimental: 5270 Age: 12–14 Schools: from five school districts	*Understanding:* Test instruments from the *Iowa Assessment Package* that gathers data in six domains including concept (i.e., understanding) and attitude *Attitude:* See above Data also explored for gender and ability effects Data on understanding gathered postintervention Data on attitude gathered pre- and postintervention

Zoller et al. (1990)	A G	*N* = 473 Control: 276 Experimental: 101 Age: 16–17 Schools: 6	Four views on science–technology–society (VOSTS) items Data also explored for gender effects Data gathered postintervention
Zoller et al. (1991)	A	577 students Control: 255 Experimental: 302 Age: 16–17 Schools: 6	Six VOSTS items. Data also explored for gender effects. Data gathered postintervention, but also gathered from 96 students only part way through the intervention to give a form of preintervention data

[1] U = Understanding, A = attitude, g = Gender, Ab = abilility.

questions. The standard assessment items more closely resemble questions on more conventional examination papers. 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. Such an implication is supported by the findings of the Wierstra (1984) study, which reports that students taking the context-based course, *PLON*, performed better in the *PLON*-style questions on the understanding instrument, and students following more conventional courses performed better on more traditional style questions.

How Big Are the Effects?

There has been considerable emphasis on "effect sizes" in recent research literature on evaluation studies. *Effect size* is simply a way of quantifying the size of the difference between two groups by looking at the average improvement in students' marks in assessment tests and comparing it with the range of scores (the standard deviation) found for typical groups of students on the same tests. Effect sizes tend to be described as "small" if less than 0.2, and "large" if greater than 0.4 (e.g., Cohen, 1969). Typically, educational interventions tend to have small effect sizes.

Of the four studies that report improved understanding, none report effect sizes per se. Two studies presented statistical analysis (Winther & Volk, 1994; Yager & Weld, 1999) in the form of t-tests. Results were significant at the 0.05 level in the Winther and Volk study, and at the 0.01 level in the Yager and Weld study. Effect sizes can, however, be calculated from the data supplied in some of the studies. In the Winther and Volk intervention, the effect size was of 0.63; and, for the Yager and Weld intervention, the effect size was 1.52. In this last case, the effect size is exceptional for an educational intervention. On the basis of the evidence resented in the study, this does appear to have been a particularly successful intervention in terms of gains in understanding of science. However, it is worth noting that the instrument used to test levels of understanding was developed by the researchers themselves as part of an ongoing research and development program on STS education, and the issues concerning style of assessment items identified in the Barber (2000) and Wierstra (1984) studies might also be of relevance here.

Because the Banks (1997) study made use of diagnostic questions with open responses and the Tsai (2000) study gathered data through interviews, statistical analysis is inappropriate in these cases.

Taken together, the findings on understanding of science provide strong 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.

A literature review into the effect of the use of socioscientific issues on reasoning (Sadler, 2004) concludes that socio-scientific issues, i.e. everyday contexts, lend themselves to integrating school science into personal science, thus helping the development of understanding of science concepts. Sadler suggests that, for such integration, either local issues should be used as contexts or teaching needs to include an explicit attempt to link more general socioscientific issues to the students' own experiences. Along the same lines, evaluations of interventions using historical case studies as "contexts" (for instance, Klopfer & Cooley, 1963; Irwin, 2000) emphasize that an increase in students' understanding of contemporary science content depends on the extent to which they take the case study as a personal challenge requiring a creative response, taking account of the historical limitations.

This review provides some evidence to suggest that performance on assessment items is linked to the nature of the items used, i.e. students following context-based/STS courses perform better on context-based questions than on more conventional questions. Studies on other context-based teaching strategies (for instance, Cho, 2002; Lubben, Bennett, Hogarth, & Robinson, 2002) conclude that only a departure from a formal teaching style allows students to gain the full benefits from a context-based approach in terms of the development of their understanding of science concepts.

THE EVIDENCE ON ATTITUDES TO SCHOOL SCIENCE AND SCIENCE

The evidence on attitudes to school science and to science comes from the findings of nine studies, one of high quality, three of medium high quality, and five of medium quality (see Table 2 for details).

By far the most common approach to gathering data on attitude was the use of inventories involving agreement/disagreement scales (Likert-type questionnaires). These were used in six studies (Barber, 2000; Ben-Zvi, 1999; Smith & Matthews, 2000; Wierstra, 1984; Wierstra & Wubbels, 1994; Yager & Weld, 1999). In all but one of these cases (Yager & Weld, 1999), the instruments were developed by the researchers specifically for the study.

In the remaining three studies, Key (1998) gathered her data via an open questionnaire, and both studies by Zoller, Donn, Wild, and Beckett (1991) and Zoller et al. (1990) used selected items from the views on science–technology–society (VOSTS) instrument developed in Canada (Aikenhead & Ryan, 1992).

Seven of the nine studies report evidence that indicates context-based/STS approaches improve attitudes to school science (or aspects of school science) and/or science more generally (Barber, 2000; Key, 1998; Smith & Matthews, 2000; Wierstra, 1984; Yager & Weld, 1999; Zoller et al. 1990, 1991). On the basis of information provided in the study reports, three of the seven studies focused on attitudes to school science, or individual science subjects: Barber (2000) on school chemistry, Smith and Matthews (2000) on school science, and Wierstra (1984) on school physics. Smith and Matthews also report data on attitudes to science. Yager and Weld (1999) do not provide details of the focus of their instrument. Key (1998) and Zoller et al. (1990, 1991) have a more limited focus to their studies, with Key reporting on attitudes to chemical industry, and Zoller et al. using four items from a much larger, extensively validated, databank of items, the VOSTS items to explore student' views on specific aspects of science in relation to society. Thus, the strength of the evidence from the Key and Zoller studies is more limited.

Of the nine studies, three present data that have been subjected to statistical analysis. Barber (2000) reports $\chi^2 = 4.94$ significant at the 0.05 level. Wierstra (1984) reports the outcome of t-tests as being significant at the 0.05 level, though insufficient data are presented to calculate effects sizes. Yager and Weld (1999) report analysis of covariance data that indicates that their findings are significant at the 0.01 level. Calculation of the effect size from their data gives a figure of 0.67.

The remainder of the studies either employed simple descriptive statistics or gathered data for which statistical analysis was inappropriate.

No improvement in attitudes was reported in the Ben-Zvi (1999) study, and the Wierstra and Wubbels (1994) study reported less positive attitudes as a result of their intervention.

Three studies also collected data relating to subject choices beyond the compulsory period and/or career intentions, as these are seen as important indicators of attitude to the subject. Here, the evidence is mixed. In two cases, increases in numbers electing to study science subjects were reported. Barber (2000) found significantly more students taking the context-based course *Salters Advanced Chemistry* elected to study chemistry at the tertiary level. Smith and Matthews (2000) report significant increases in numbers choosing physics and biology in students taking the STS program, with numbers doubling in both subjects. (No STS option was offered to chemistry students, and there was no increase in numbers choosing chemistry.) In contrast, Ramsden (1997), though not reporting on attitudes in any detail, found that more positive responses to a context-based approach in lessons were not translated into increased interest in careers involving science.

It is worth noting that one feature which distinguishes the Barber (2000) study and the Smith and Matthews (2000) study from the Ramsden (1997) study is that, in the latter case, the author was not the teacher of the students from whom the data were collected. This points to the possibility of individual teacher effects exerting a strong influence on students.

Taken together, the findings on attitudes to school science and science provide very strong evidence that context-based/STS approaches foster more positive attitudes to school science than conventional courses. There is more limited evidence to suggest context-based/STS approaches foster more positive attitudes to science more generally than conventional courses. There is mixed evidence on the impact of context-based/STS approaches on subject and career choices.

Results from nonexperimental studies confirm these findings. For instance, Lyons (2006) reports in a three-country comparative study (Australia, Sweden, and the United Kingdom) that attitudes to science learning are strongly influenced by students' perceptions of "how it is related to the real world and technology and the future" (p. 599). Häussler and Hoffmann (2000) explored the types of contexts that generate students' interest in learning of physics at German secondary schools. They conclude that contexts from the socioscientific and emotional sphere improve attitudes to physics learning more than contexts emphasizing practical activities, intellectual development, or preparation for future work. In a large-scale attitude survey of students in Scotland taking courses which are context based to different degrees, Reid and Skryabina (2002) find that throughout secondary schooling context-based learning approaches are related to a higher interest in further science training and science-based careers. However, at preuniversity level further study and career choice is linked less prominently to context-based approaches.

THE EVIDENCE ON GENDER AND ABILITY EFFECTS

The evidence on gender effects comes from the findings of five studies, one of high quality (see Table 2 for details). Given the small number of medium quality or better studies that explored gender effects, these findings are presented only briefly. Only one study (Yager & Weld, 1999) presented detailed statistical data of direct relevance to the review. The evidence on ability effects is very limited, coming only from one study (Yager & Weld).

Three studies suggest that gender differences in attitudes are reduced through the use of a context-based approach (Smith & Matthews, 2000; Wierstra, 1984; Yager & Weld, 1999). The Yager and Weld study provides statistical data in the form of analysis of covariance, with results significant at the 0.01 level. There was also evidence to suggest that girls in classes using a context-based/STS approach held more positive attitudes to science than their female peers in classes using a conventional approach (Smith & Matthews, 2000; Yager & Weld, 1999). The Yager and Weld study provides statistical data in the form of analysis of covariance, with results significant at the 0.01 level. There was also some evidence from this study to suggest that girls following context-based/STS courses were more positive than their peers following conventional courses to pursuing careers involving science.

Taken together, these findings suggest that there is moderate evidence to indicate that context-based/STS approaches promote more positive attitudes to science in both girls and boys and reduce the gender differences in attitudes.

The Yager and Weld study indicated that lower ability pupils in classes using a context-based/STS approach developed a better conceptual understanding of science and held significantly more positive attitudes to science than their lower ability peers taking conventional courses. They also developed a better conceptual understanding of science and held significantly more positive attitudes toward science than their higher ability peers in the same classes. All results were significant at the 0.01 level.

ISSUES IN RESEARCH INTO THE EFFECTS OF CONTEXT-BASED/STS APPROACHES

Evaluation Designs

This paper has focused on evaluations with experimental designs. It is of note that a sizeable majority of the evaluation studies identified did *not* adopt experimental designs and only gathered data from students experiencing the intervention. Such a feature of work in the area might be attributed to one or more of the following: lack of funding to support detailed experimental evaluation, difficulty of identifying appropriate control groups, practical constraints in setting up experimental studies (e.g., negotiating access to schools), or not seeing the gathering of comparative data as a necessary feature of an evaluation.

There is currently considerable interest in evaluations of this nature, particularly where they have involved what some see as the "gold standard" (Torgerson & Torgerson, 2001) of research design, the randomized control trial, or RCT. RCTs are seen as providing the strongest evidence of "what works" (Oakley, 2000). They are very commonly used in medical research and involve the random allocation of subjects to control and experimental groups to evaluate the effects of an intervention.

It is interesting that only one of the studies in this report (Tsai, 2000) 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 can rarely be made by researchers, but, instead, may depend on decisions by policy makers or groups of practitioners such a school departments. This means that it is very 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 studies in this review, the sampling often had an opportunistic dimension 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 make it necessary to gather data from intact classes and raise issues to do with the construction of matched samples for control and experiential groups. In some cases, as can be seen in Table 4, the researchers made the decision not to use matched samples but report all the data they gathered. In other cases, steps were taken to control some of the variables in constructing samples. For example, in the Ramsden (1997) study, data were collected from intact classes, but the control and experimental groups that formed the basis of the analysis were a subset of the intact classes, matched in terms of ability on the basis of their predicted grades in their forthcoming external examination. Taking steps to match control and experimental groups as closely as possible enhances the validity of any claims made about effects.

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 educational evaluation. Before it is possible to decide "what works," it is necessary to decide what "working" means. This can be illustrated with reference to the Barber (2000) study. As noted earlier in this paper, Barber's findings showed that 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.

It seems perfectly reasonable to suggest that, if the aims of an intervention are different, the way that it should be assessed it also likely to be different. The Barber study established that students in both groups achieved similar grades in their final examinations. (All courses and examinations are assessed by an external group to ensure standards are comparable.) Thus if "working" 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." Most context-based and STS interventions involve a shift in the intended outcomes, and the old and the new cannot therefore be directly compared. This implies that aspects of such interventions have to be evaluated against their declared aims, and not by comparisons with another approach.

Standardization of Instruments

The variety of instruments used in the studies to gather data, particularly attitudinal data, means it is not feasible to make direct comparisons between studies or undertake meta-analysis of data. This raises the question of how feasible it might be to make use of standardized instruments in the evaluation of context-based approaches. While in principle this would appear to have some merit, in practice, a number of problems emerge, particularly when trying to make international comparisons: 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 militate against the validity of using some form of crossnational measure of scientific understanding. There does, however, appear to be more scope for the use of standardized diagnostic questions (i.e., those that require students to provide explanations of their answers) across a range of settings and countries.

In theory, the development of a standardized instrument to measure attitudes appears to offer more scope. However, as the literature of the last three decades reveals (e.g., from Gardner, 1975; Ramsden, 1988; Schibecci, 1984), the design of reliable and valid attitudinal instruments remains problematic. Whilst detailed analysis of attitude studies (e.g., Ramsden, 1998) reveals that fixed-response and scaling techniques are a consistent feature of attitude research, the design of individual instruments often reveals weaknesses in item structure and/or analysis. It is beyond the scope of this paper to examine in detail the range of instruments used in the study reviewed, but, in order to be included in the review, all had to provide details of satisfactory checks on reliability and validity. Nonetheless, it is possible to identify potential weaknesses in design. For example, attitudinal instruments often contain a number of statements to which respondents have to indicate their view on an agree/disagree scale. Instruments used in the studies reported here, in common with a number of other attitudinal instruments, contained statements not indicative of attitude. For example, a positive response to a statement such as *My family would like me to study chemistry* does not say anything about the respondent's own attitude, in contrast to a positive response to a statement such as *If I had a choice, I would choose to study chemistry.* Any new attitudinal instrument would need to be done with careful reference to the literature on instrument design and validation.

Who Collects the Data and for What Purposes

Two of the features noted when summarizing the studies for the review were the relationship of the study author(s) to the interventions being evaluated and the purposes for which the data were being collected. It was very noticeable that this information was difficult to identify in the study reports and, in almost all cases, had to be drawn by inference.

Banks (1997), Barber (2000), and Barker (in Barker & Millar, 1996) were all users of the intervention and collected their data for personal interest as part of their studies for a higher degree. Several of the authors of the study reports were also involved in the development of the materials (Ben-Zvi, 1999; Lubben et al., 1997; Ramsden, 1997; Tsai, 2000; Wierstra, 1984; Wierstra & Wubbels, 1994; Yager & Weld, 1999), and this also seemed likely to be the case for Rubba. Smith (in Smith & Matthews, 2000) both developed the intervention and collected the data for a higher degree.

The involvement of the developers in the evaluation raises a number of issues. Cast positively, it indicates they are eager to explore the effects of their interventions. However, it also reflects the fact that most, if not all, of the funding for many interventions is associated with the development, rather than evaluation. While there are signs that this is now changing (e.g. two large-scale funders in the United Kingdom now set aside a portion of the funding for external evaluation), a lack of funding for detailed systematic evaluation does impose severe limitations on the nature and scope of any data that can be collected. Moreover, the involvement of the developers in the evaluation does raise ethical issues about introducing possible bias into the findings, as it could be argued that developers have a vested interest in demonstrating that their intervention has been successful. However, the studies included in this paper all appeared to take steps to minimize the effects of any bias.

The Nature of the Resources

The information in the study reports 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 study reports that the terms "context-based approaches" and "STS approaches" can be interpreted quite broadly. Examination of the intervention resources that were available shows the use of contexts that

- are relevant to students' lives and interests at present;
- are relevant to situations students may encounter at some point in their lives;
- relate to technological developments and artifacts likely to be of interest to students;
- are relevant to students' possible future careers;

and at advanced levels of study:

- link to recent scientific research and innovations;
- link to industry.

This diversity of contexts suggests that some caution is needed in interpreting the findings of this review, as it seems difficult to imagine that all contexts have the same effects on all students. However, this caveat does need to be set against a background of the consistency of the evidence yielded by the studies taken as a whole.

Possible Further Areas of Research

The overview, or systematic map, is of interest both in terms of what is included and what is absent, as the latter may well point to areas where further work might be fruitful.

There would certainly appear to be benefits in supplementing the existing data on groups of student who, traditionally, have been particularly alienated by science: girls and lower ability students. There is a noticeable absence of studies on students from ethnic minority groups.

The broad interpretation of the terms "context-based approaches" and "STS approaches" points to the value of exploring the effects of particular types of context. Linked to this, context-based/STS approaches incorporate a wide range of activities, some of which are not traditionally associated with science teaching, and the effects of particular types of activity, for example, small-group discussions, would be worth exploring in more detail.

CONCLUSIONS

What has been learned from the process of undertaking the review? The authors believe that systematic review methods have considerable potential as a tool for synthesizing research. It is the case that the impetus for introducing systematic reviews came from criticism of nonexperimental approaches to research, and, as such, the systematic reviews have become rather more closely linked than is desirable with experimental designs. However, the approach can be adapted to synthesize findings from a range of studies and, as such, has much to offer. With respect to the substance of the findings in relation to the effects of context-based/STS approaches to the teaching of science, a number of issues have emerged. Of particular interest is the matter of the involvement of the developers of an intervention in the evaluation of its effects, which poses a question over the validity of the findings. The detailed scrutiny of the studies involved in undertaking a systematic review has reassured us that those studies which are judged to be of sufficiently good quality to be included in the review have taken sufficient steps to ensure that both the instruments used and the analysis techniques employed are reliable and valid, and that all reasonable steps have been taken to eliminate researcher bias. It is the case, however, that many of the teachers in schools participating in interventions do so because what is being proposed seems attractive and they can see potential benefits to their students. This, inevitably, is likely to result in the interventions being more successful than if they were being used by people less persuaded of their benefits. This is a feature of some curriculum intervention which is difficult to avoid—indeed, it could be argued that it benefits the process of innovation and change. Despite the caveats, the authors believe that the evidence presented in this review provides reliable and valid evidence to support the use of contexts as a starting point in science teaching: there are no drawbacks in the development of understanding of science, and considerable benefits in terms of attitudes to school science.

REFERENCES

Aikenhead, G. (1994). What is STS teaching? In J. Solomon & G. Aikenhead (Eds.), STS education: International perspectives on reform. New York: Teachers College Press.

Aikenhead, G., & Ryan, A. (1992). The development of a new instrument: Views on science–technology–society (VOSTS). Science Education, 76(4), 477–491.

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. (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. Retrieved August 24, 2006, from http: eppi.ioe.ac.uk/EPPI.

Black, P. (1998). Testing: Friend or foe? Theory and practice of assessment and testing. London: Falmer Press.

Borman, G. (2002). Experiments for educational evaluation and improvement. Peabody Journal of Education, 77, 7–27.

Brown, A. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. Journal of the Learning Sciences, 2(2), 141–178.

Burkhardt, H., & Schoenfeld, A. (2003). Improving educational research: Towards a more useful, more influential, and better-funded enterprise. Educational Researcher, 32(9), 3–14.

Byrne, D. (2005). Complexity, configurations and cases. Theory, Culture and Society, 22(5), 95–111.

Cho, J. (2002). The development of an alternative in-service programme for Korean science teachers with an emphasis on science–technology–society. International Journal of Science Education, 24(10), 1021–1035.

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

Cohen, M., Manion, L., & Morrison, K. (2000). Research methods in education. London: Routledge.

Elliot, J. (2001). Making evidence-based practice educational. British Educational Research Journal, 27(5), 555–574.

Gardner, P. (1975). Attitudes to science: A review. Studies in Science Education, 2, 1–41.

Gorard, S., & Taylor, C. (2004). Combining methods in educational and social research. Berkshire: Open University Press.

Hammersley, M. (2001). On ‘systematic’ reviews of research literature: A ‘narrative’ response to Evans and Benefield. British Education Research Journal, 27(5), 543–554.

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

Häussler, P., & Hoffmann, L. (2000). A curricular frame for physics education: Development, comparison with students’ interests, and impact on students’ achievement and self-concept. Science Education, 84, 689–705.

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

Irwin, A. (2000). Historical case studies: Teaching the nature of science in context. Science Education, 84(5), 5–26.

Kaestle, C. (1993). The awful reputation of educational research. Educational Researcher, 22(1), 23–31.

Kelly, A. (1986). The development of children’s attitudes to science. European Journal of Science Education, 8(4), 399–412.

Klopfer, L., & Cooley, W. (1963). The *History of Science Cases* for high schools in the development of student understanding of science and scientists. Journal of Research in Science Teaching, 1(1), 33–47.

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 students. In: Research Evidence in Education Library. London: EPPI-Centre, Social Science Research Unit, Institute of Education. Retrieved August 24, 2006, from eppi.ioe.ac.uk/EPPI.

Lubben, F., Campbell, B., & Dlamini, B. (1996). Contextualising science teaching in Swaziland: Some student reactions. International Journal of Science Education, 18(3), 311–320.

Lyons, T. (2006). Different countries, same science classes: Students’ experiences of school science in their own words. International Journal of Science Education, 28(6), 591–613.

Messick, S. (1995). Validity for psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749.

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

Oppenheim, A. N. (1992). Questionnaire design, interviewing and attitude measurement. London: Pinter Publishers.

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. International Journal of Science Education, 25(9), 1049–1079.

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), 206–219.

Ramsden, J. (1998). Mission impossible: Can anything be done about attitudes to science? International Journal of Science Education, 20(2), 125–137.

Reid, N., & Skryabina, E. (2002). Attitudes towards physics. Research in Science and Technology Education, 20(10), 67–81.

Rosnow, R., & Rosenthal, R. (1996). Computing contrasts, effect sizes and counternulls on other people’s published data: General procedures for research consumers. Psychological Methods, 1, 331–340.

Sadler, T. (2004). Informal reasoning regarding socio-scientific issues: A critical review of research. Journal of Research in Science Teaching, 41(5), 513–536.

Schibeci, R. (1984). Attitudes to science: An update. Studies in Science Education, 11, 26–59.

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

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

Torgerson, C., & Torgerson, D. (2001). The need for randomised controlled trials in educational research. British Journal of Educational Studies, 49(3), 316–328.

Ukoumunne, O., Gulliford, M., Chinn, S., Sterne, J., Burney, P., & Donner, A. (1999). Methods in health service research: Evaluation of health interventions at area and organisation level. British Medical Journal, 319, 376–379.

U.S. Department of Education. (2003). Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide. Washington, DC: Institute of Education Sciences.

Vulliamy, G. (2004). The impact of globalisation on qualitative research in comparative and international education. Compare, 34(3), 261–284.

REFERENCES FOR THE STUDIES INCLUDED IN THE REVIEW

Banks, P. (1997). Students' understanding of chemical equilibrium. Unpublished MA thesis, University of York, UK.

Barber, M. (2000). A comparison of NEAB and Salters' A-level chemistry: Students' views and achievements. Unpublished MA thesis, University of York, UK.

Barker, V., & Millar, R. (1996). Differences between Salters' and traditional A-level chemistry students' understanding of basic chemical ideas. York, UK: University of York.

Ben-Zvi, R. (1999). Non-science oriented students and the second law of thermodynamics. International Journal of Science Education, 21(12), 1251–1267.

Key, M-B. (1998). Students' perceptions of chemical industry; influences of course syllabi, teachers, firsthand experience. York, UK: University of York.

Lubben, F., Campbell, B., & Dlamini, B. (1997). Achievement of Swazi students learning science through everyday technology. Journal of the Southern African Association for Research in Mathematics and Science Education, 1(1), 26–40.

Ramsden, J. M. (1997). How does a context-based approach influence understanding of key chemical ideas at 16+? International Journal of Science Education, 19(6), 697–710.

Rubba, P. A., McGuyer, M., & Wahlund, T. M. (1991). The effects of infusing STS vignettes into the genetics unit of biology on learner outcomes in STS and genetics: A report of two investigations. Journal of Research in Science Teaching, 28(6), 537–552.

Smith, L. A., & Bitner, B. L. (1993). Comparison of formal operations: Students enrolled in ChemCom versus a traditional chemistry course. Paper presented at the Annual Meeting of the National Science Teachers Association, April 1993, Kansas City, MO.

Smith, G., & Matthews, P. (2000). Science, technology and society in transition year: A pilot study. Irish Educational Studies, 19, 107–119.

Tsai, C-C. (2000). The effects of STS-oriented instructions on female tenth graders' cognitive structure outcomes and the role of student scientific epistemological beliefs. International Journal of Science Education, 22(10), 1099–1115.

Wierstra, R. F. A. (1984). A study on classroom environment and on cognitive and affective outcomes of the PLON-curriculum. Studies in Educational Evaluation, 10(3), 273–282.

Wierstra, R. F. A., & Wubbels, T. (1994). Student perception and appraisal of the learning environment: Core concepts in the evaluation of the PLON physics curriculum. Studies in Educational Evaluation, 20(4), 437–455.

Winther, A. A., & Volk, T. L. (1994). Comparing achievement of inner-city high school students in traditional versus STS-based chemistry courses. Journal of Chemical Education, 71(6), 501–505.

Yager, R. E., & Weld, J. D. (1999). Scope, sequence and coordination: The Iowa Project, a national reform effort in the USA. International Journal of Science Education, 21(2), 169–194.

Zoller, U., Donn, S., Wild, R., & Beckett, P. (1991). Students' versus their teachers' beliefs and positions on science/technology/society-oriented issues. International Journal of Science Education, 13(1), 25–36.

Zoller, U., Ebenezer, J. V., Morley, K., Paras, S., Sandberg, V., West, C. et al. (1990). Goal attainment in science–technology–society (S/T/S) education and reality: The case of British Columbia. Science Education, 74(1), 19–36.