REVIEW

August 2005

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

Review conducted by the TTA-supported Science Review Group

The EPPI-Centre is part of the Social Science Research Unit, Institute of Education, University of London

AUTHORS/REVIEW GROUP

Dr Judith Bennett	University of York
Dr Sylvia Hogarth	University of York
Fred Lubben	University of York
Alison Robinson	University of York

Address for correspondence
Fred Lubben
Department of Educational Studies
University of York
York YO10 5DD

Tel: 01904 432560
Email: fel1@york.ac.uk

CONFLICTS OF INTEREST

Two members of the Review Group, Judith Bennett and Fred Lubben, have been involved in the development of school science materials which emphasise context-based approaches to the teaching of science. All members of the Review Group have been involved in evaluating context-based courses and have published in this area. In fact, one of the authors of the Review (Bennett) is the author of two of the fourteen studies included in the in-depth review (Ramsden, 1992, 1997). In addition, the Review Group members are members of the University of York Science Education Group (UYSEG), which is currently developing a new secondary level science course, *21st Century Science*, aimed at developing pupils' scientific literacy. However, none of the Review Group is directly involved in the development of this course.

ACKNOWLEDGEMENTS

The Science Review Group is one of a number of teams undertaking systematic reviews to support the evidence base of Initial Teacher Training (ITT) staff in England.

The Review Group is grateful for the funding from the Teacher Training Agency (TTA) which enabled this review to be carried out. The Review Group also acknowledges the financial support via core institutional funding from the Higher Educational Funding Council for England (HEFCE). The Review Group is grateful to the Evidence for Policy and Practice Co-ordinating Centre (EPPI-Centre) at the Institute of Education, University of London, for training and support for the review, and for permission to use their systematic review tools.

LIST OF ABBREVIATIONS

BEI	British Education Index
CAIMI	Children’s Academic Intrinsic Motivation Inventory
ChemCom	Chemistry in the Community
ChiK	Chemistry in Context (German)
DfES	Department for Education and Skills
EPPI-Centre	Evidence for Policy and Practice Information and Co-ordinating Centre
ERIC	Education Resources Information Centre
GALT	Group assessment for logical thinking
GCSE	General Certificate for Secondary Education
IAP	Iowa Assessment Package
ILL	Inter-library loan
ITT	Initial teacher training
PISA	Programme for International Student Assessment
PLON	Physics Curriculum Development Project (Dutch)
QCA	Qualifications and Curriculum Authority
RCT	Randomised controlled trial
SATIS	Science and technology in society
SAT	School assessment test
SD	Standard deviation
SS&C	Scope, Sequence & Continuity
SSCI	Social Science Citation Index
STS	Science, Technology and Society
TIMMS	Third International Mathematics and Science Survey
TSCS	Tennessee Self-Concept Scales
TTA	Teacher Training Agency
VOSTS	Views On Science Technology and Society
WoE	Weight of evidence

This report should be cited as: Lubben F, Bennett J, Hogarth S, Robinson A (2005) 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, University of London.

TABLE OF CONTENTS

SUMMARY

Background

This review focuses on context-based approaches which promote links between science, technology and society (STS) in the teaching of science in secondary schools. The reasons for selecting these teaching approaches for this review are as follows:

- Courses adopting context-based and STS approaches have attracted national and international attention as they are seen to have an important role to play in developing pupils' scientific literacy.
- Context-based and STS approaches are strongly advocated in the *National Curriculum for Science* in England and in Wales, and in the *Qualifications and Curriculum Authority (QCA) Scheme of Work for Key Stage 3.*
- Context-based and STS approaches have been advocated for a number of years on ITT courses as a means by which teachers might motivate their pupils.
- The current move towards developing school science courses which emphasise scientific literacy will have significant implications for science ITT courses. The overlap between the aims of context-based/STS courses and scientific literacy courses, together with the approaches they advocate, makes it highly desirable to establish the strength and nature of the evidence base for the claims made for such approaches.

This review builds on work undertaken for an earlier systematic review report (Bennett *et al.*, 2003) on the effects on pupils of teaching approaches which emphasise placing science in context and promote links between science, technology and society (STS). This earlier 'base' review included studies which had explored effects on both understanding of science ideas and attitude to science. The review reported here looks in detail at two groups of pupils traditionally alienated by conventional approaches to science teaching: (i) girls and (ii) low-ability pupils.

Aims

The review has two principal aims:

- to explore the effects of context-based and STS teaching approaches on boys and girls, and on lower-ability pupils
- to inform the evidence base on which initial teacher training (ITT) courses draw in relation to the above teaching approaches

Review questions

The overall review question is identical to that of the earlier 'base' review:

What evidence is there that context-based and STS teaching approaches improve the understanding of science ideas and the attitudes to science of 11- to 18-year-old pupils, and what are the implications of the evidence for initial teacher training courses?

The focus on understanding encompasses science concepts, ideas about the nature of science, and scientific method. The focus on attitude encompasses attitude towards science, attitude towards school science, motivation to learn, interest in science activities, and career intentions.

From the earlier review, it became apparent that a number of studies made reference to other science-related abilities, such as the development of investigative skills, manipulative skills, communication skills, problem-solving skills, and decision-making skills. While this area was not the main focus of the review, such effects were noted when studies were categorised (keyworded) to produce an overview (the systematic map). It is recognised that, although some of these skills are not science-specific, there is an expectation that they will be applied in science contexts. It is further recognised that there are areas of debate over the interpretation of some of these terms (e.g. attitude, problem-solving skills) and how they might be measured.

In producing the systematic map of the area, it became apparent that the number of studies involved required a narrower focus to create a subset for the in-depth review. It was therefore decided to focus the in-depth review on studies of effects on understanding and attitudes of pupils who followed *whole courses* using context-based approaches.

The review questions for the two in-depth review areas are as follows:

What is the evidence from evaluative studies of the effect of context-based or STS courses on the attitude to science and/or the understanding of science ideas of boys and girls in the 11 to 16 age range?

What is the evidence from evaluative studies of the effect of context-based or STS courses on the attitude to science and/or the understanding of science ideas of lower-ability pupils in the 11 to 16 age range?

Methods

The review methods are those developed by the EPPI-Centre for systematic reviews of educational research literature. Such a review has four main phases:

- *Searching and screening:* developing criteria by which studies are to be included or excluded in the review, searching (through electronic databases and by hand) for studies which appear to meet these criteria, and then screening the studies to see if they meet the inclusion criteria
- *Keywording and generating the systematic map:* coding each of the included studies against a pre-agreed list of characteristics which is then used to

generate a systematic map of the area where studies are grouped according to their chief characteristics

- *In-depth review and data extraction:* summarising and evaluating the contents of studies according to pre-agreed categories
- *Synthesis:* providing an overview of the quality and relevance across the studies in the in-depth review and compiling the weighted findings of the collective studies

The review presented in this report makes use of the results of the searching and screening phase completed as part of the 'base' review. No new studies have been added to the systematic map providing an overview of the characteristics of the studies identified, but an additional paper that reported on one of the included studies was obtained and the map has been modified to reflect findings of 61 studies, rather than 67 papers; six pairs of papers reported on one study each.

The remaining two phases – data-extraction and synthesis – are specific to the new in-depth review questions.

Results

The systematic map of 61 studies reveals a number of characteristics of research on the effect of context-based and STS approaches to teaching science, as summarised below.

- One in three of the studies report work that has taken place in the USA, and one in four report work in the UK.
- Past and currently active research groups in this area are based at the Universities of British Columbia (Canada), Cambridge (UK), Iowa (USA), Utrecht (the Netherlands) and York (UK).
- Two in three studies were undertaken with pupils aged 11–16, the remainder with pupils between the ages of 17 and 20.
- More than half the studies reported on context-based/STS interventions in science, the remainder being equally split between chemistry and physics. Very little research has been done on context-based/STS approaches in the teaching of biology.
- Almost two out of three studies evaluated interventions using context-based approaches, the majority of which concern full courses. The remaining third of the studies evaluated interventions using STS approaches, equally divided between STS-enrichment activities and full STS courses.
- About 60% of the studies are naturally occurring evaluations, and the remaining 40% are researcher-manipulated.
- The most popular techniques for gathering data are test results, questionnaires, (dis)agreement scales and interviews. The first is over-represented amongst researcher-manipulated evaluations, the last amongst naturally occurring evaluations.

Fourteen studies were included in the in-depth review, which focused on the effects of context-based or STS courses on the attitude to science and/or the understanding of science ideas of boys and girls, and lower-ability pupils, in the 11 to 16 age range.

The consolidated evidence from this in-depth review draws primarily on the findings from studies weighted as *high*, *medium-high* and, to a lesser extent, as *medium*, as summarised above. Findings from studies weighted as *medium-low* are only considered if these corroborate findings of studies with a higher weight of evidence.

The small number of studies considered for the in-depth review are of variable quality. Therefore many of the findings have, on purpose, been cast in tentative terms because of their narrow evidence base. For that reason the findings below have been reported under two headings: that is, those supported by *reasonable evidence*, and those supported by *some evidence*. No findings are claimed to be based on *strong evidence*.

The review suggests that there is ***reasonable*** evidence of the following:

- *Girls in classes using a context-based/STS approach held significantly more positive attitudes to science than their female peers in classes using a traditional approach* (based on mutually supportive evidence from one study rated high, one rated medium-high, and one study rated medium).

- *Similarly, boys in classes using a context-based/STS approach held significantly more positive attitudes to science than their male peers in classes using a traditional approach* (based on mutually supportive evidence from the same set of studies as above).

- *A context-based/STS approach to teaching science narrowed the gap between boys and girls in their attitude to science* (based on mutually supportive evidence from one study rated high, two rated medium-high, and two studies rated medium; supplementary support from two other studies).

- *In cases when boys enjoyed the materials significantly more than girls, this was due to the nature of the practical work in the unit; in cases when girls enjoyed context-based materials significantly more than boys, this was because of the non-practical activities in the unit* (based on mutually supportive evidence from two studies rated medium-high, and one study rated medium).

The review suggests there is ***some*** evidence of the following:

- *Boys and girls in classes using a context-based approach significantly more often perceived a close link between science, technology and society than their gender peers in traditional classes; there were slight gender-related differences in the way science was linked to technology and society* (based on evidence from one study rated medium).

- *Boys and girls in classes using a context-based/STS approach showed significantly better conceptual understanding of science than their gender peers in classes using a traditional approach* (based on evidence from one study rated high).

- *Girls in classes using a context-based/STS approach developed a significantly more positive attitude towards taking a science career compared with boys in* *these classes* (based on contradicting evidence from one study rated high, supporting this conclusion, and one study rated medium-high, concluding that girls and boys have the same, both very positive, attitude to pursuing a science career).

- *Girls in classes using a context-based/STS approach showed equal conceptual understanding of science as male peers in the same classes* (based on contradictory evidence from one study rated high, supporting the conclusion, and one study rated medium, concluding the boys using a context-based/STS approach significantly outperform girls using the same approach).

- *Lower-ability pupils in classes using a context-based/STS approach held significantly more positive attitudes to science than lower-ability pupils in classes using a traditional approach* (based on evidence from one study rated high; supplementary support from one other study).

- *Lower-ability pupils in classes using a context-based/STS approach developed significantly more positive attitudes towards science than high-ability peers in the same classes* (based on evidence from one study rated high; supplementary support from one other study).

- *Lower-ability pupils in classes using a context-based/STS approach showed significantly better conceptual understanding of science than their lower ability peers in classes using a traditional approach* (based on mutually supportive evidence from one study rated high and one study rated medium).

- *Lower-ability pupils in classes using a context-based/STS approach showed higher gain in conceptual understanding of science than high-ability peers in the same classes* (based on evidence from one study rated high).

Conclusions

Strengths and limitations

Strengths of the review

The review has a number of strengths as follows:

- The review focus is highly topical. The current concern about low uptake of science studies and careers by girls, and about under-achievement of boys has reignited interest in gender issues. Further evidence of the topicality comes from the range of countries in which studies have been undertaken.

- The review has established that there is consistency in the research approaches adopted by those researching the effect of context-based or STS approaches to science teaching on pupils' understanding of science ideas or on their attitude to science. Such approaches make use of a pre/post quasi-experimental research design and generate quantitative data. For measuring understanding of science ideas, standardised achievement tests are used to compare the effects of different curricula. For measuring attitude to science, a large variety of instruments are used, using Likert- or Osgood-type scales.

- Quality-assurance results were high for all stages of the review.

Limitations of the review

The review has two main limitations:

- There was a scarcity of studies that focused on the effects of context-based or STS approaches for boys and girls, or for lower-ability pupils, as an independent variable. Only seven studies were judged to be of reasonable quality with respect to the review question (that is, an overall weight of evidence of medium or higher).
- Although the studies in the in-depth review share a number of similar characteristics at the broad level, there are considerable differences at the detailed level. Thus, the synthesis could only claim *reasonable* or *some* evidence (as opposed to *strong* evidence) for any of the conclusions drawn.

Implications

The Review Group is cautious about commenting on implications of the review for policy and practice for the reasons given in the preceding section on 'Limitations'.

Implications for policy

The review has yielded reasonable evidence that both girls and boys in secondary school science classes using a context-based or STS approach develop more positive attitudes than peers following more traditional courses. Even more strongly, the review supports the conclusion that the differences between the attitudes to science of boys and girls can be decreased by learning through context-based or STS approaches.

The review therefore indicates that a policy which embraces the teaching of science through linking it to the everyday experiences of pupils is likely to have a beneficial effect on their attitudes to science. However, it should also be noted that there is a scarcity of high quality research evidence of the positive effects of such an approach for different subsets of pupils, such as those of lower ability (as indicated by this review), those from cultural minorities, second-language speakers, or those using different learning styles or sense-making strategies.

The review provides some evidence that context-based/STS approaches may foster more positive attitudes to science careers in girls in particular, although the data gathered relate only to indications of intent, not actual choices. However, there is no evidence to suggest that a policy of using context-based/STS approaches has a detrimental effect on attitudes to science careers of girls (or boys).

Implications for initial teacher training

It is highly desirable that students and tutors on initial teacher training courses should be made aware of the evidence on the effects on context-based and STS approaches on pupils' attitudes, as should those who produce resources for use on such courses. This includes the DfES (which is responsible for introducing the Key Stage 3 Strategy), QCA (which has produced an extensive and detailed

Scheme of Work for Key Stage 3), and the TTA (which specifies standards for newly qualified teachers).

It is clear from this review that context-based/STS approaches stimulate pupils' interest in science. However, there is a possible tension between using contexts as starting points and the current emphasis in many of the materials produced for science teaching which indicate lessons should start by making learning outcomes explicit to pupils. The two notions are not incompatible, but require a more sophisticated approach to lesson planning. A strategy of explicitly stating the intended lesson outcomes after a context-based lesson introduction may well avoid one of the main risks of context-based learning: that is, pupils' idea that the learning outcome is about the understanding of the context rather than the science concepts underlying the context.

Implications for practice

In order to improve attitudes to science of boys and girls, especially their interest in science lessons, teachers should use everyday interests as starting points. Several of the studies in the map provide ideas for possible contexts and teaching strategies.

In order to gain maximum benefit from the use of a context-based/STS approach for both girls and boys, teachers should make use of a variety of activities in their lessons, though should they adopt context-based or STS programmes this is going to happen as a matter of course.

Implications for secondary research

One particular area of the systematic map would benefit from further exploration. This relates to studies which report on the effects on understanding of science ideas. Present knowledge in this area, synthesised through previous systematic reviews, is seriously limited. Currently, only those studies which have explored understanding and attitude, or understanding in relation to gender or ability, have formed the basis of in-depth reviews. A broadening of the evidence-base using the existing map represents an efficient use of resources.

In the light of the effectiveness of context-based/STS approaches in improving pupils' attitudes to science, and the strong recommendation for adoption of such approaches in classrooms and ITT programmes, a consolidation of knowledge is required about teacher professional development strategies used for supporting a change practice: what characteristics of professional development strategies lead to adoption (or adaptation) of the use of context-based approaches?

What are crucial steps in the development of teacher ownership of context-based/STS approaches?

Implications for primary research

This review points to three specific areas which would benefit from primary research. Firstly, the review indicates a need for further work on the effects of context-based/STS approaches on lower-ability pupils, with much of the evidence emerging from one study. Secondly, the review also suggests that little research has been done into effects on more able pupils, again pointing to an area of further work. Thirdly, given the current concerns about boys' perceived underachievement and the indication in this review (based on limited evidence from the USA only) that context-based/STS approaches enhanced boys' understanding of science, it would be desirable to explore aspects of gender and achievement for context-based and STS courses in more detail.

More generally, both this review and the attitude and understanding review (Bennett *et al.*, 2003) have tended to treat quite sophisticated concepts as fairly simple ideas. Some of these would benefit from further unpicking. For example, the term ‘context’ can be interpreted in a wide variety of ways, and it would be useful to look at some of these ways, with a view to establishing how particular groups of pupils respond to different contexts. Such work would be particularly useful for those developing courses on public understanding of science.

By focusing on evaluations involving an experimental design, both reviews have also tended to yield evidence which relates to *what* effects context-based/STS approaches have had, rather than *why* these effects have occurred. Thus some empirical work is needed involving interviews with pupils to explore their views on particular contexts and their particular effects. Such work could also be fruitful in exploring, for example, aspects such as why more positive attitudes to science do not appear to be translated to any significant extent into a desire to pursue careers in science.

1. BACKGROUND

1.1 Aims and rationale for current review

The last twenty years have seen a number of changes take place in science teaching, of which one of the most significant has been the development of a wide range of materials which use contexts and applications as a starting point for developing understanding of scientific ideas. Such approaches are variously described as 'context-based', 'applications-led' or using 'STS' (Science-Technology-Society) links. Pupils using such materials might, for example, find out about the electromagnetic spectrum through learning about medical techniques for seeing inside the body, or explore the views of different members of a community on the impact of locating a chemical industry nearby. Examples of curriculum development drawing on such approaches can be found in materials ranging from small teaching units to whole courses, developed on local, national and international scales, and for all stages of education from primary through to tertiary.

This review builds on work undertaken for an earlier systematic review report (Bennett *et al.*, 2003) on the effects on pupils of teaching approaches which emphasise placing science in context and promote links between science, technology and society (STS). This earlier 'base' review focused on interventions which had explored effects on both understanding of science ideas and attitude to science. The review reported here looks in detail at two groups of pupils traditionally alienated by conventional approaches to science teaching: (i) girls and (ii) low-ability pupils, and explores the effects of teaching approaches which emphasise placing science in context and promote links between science, technology and society (STS) on boys and girls, and on lower-ability pupils.

The review topic falls within the broader area of *scientific literacy*, which is currently receiving considerable attention in policy-making discussions about the future of the school science curriculum. The reasons for selecting context-based approaches and approaches which promote links between science, technology and society are as follows:

- Courses adopting context-based and STS approaches have attracted national and international attention in the last two decades as they are seen to have an important role to play in developing pupils' scientific literacy.

- Context-based and STS approaches are strongly advocated in the *National Curriculum for Science* in England and Wales, and in the *Qualifications and Curriculum Authority (QCA) Scheme of Work for Key Stage 3.*

- Context-based and STS approaches have been advocated for a number of years on ITT courses as a means by which teachers might motivate their pupils.

- The current move towards developing school science courses which emphasise scientific literacy will have significant implications for science ITT courses. The overlap between the aims of context-based/STS courses and scientific literacy courses, together with the approaches they advocate, makes it highly desirable to establish the strength and nature of the evidence base for the claims made for such approaches.

1.2 Definitional and conceptual issues

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

Context-based approaches are approaches adopted in science teaching, in which contexts and applications of science are used as the *starting point* for the development of scientific ideas. This contrasts with more traditional approaches, which cover scientific ideas first and conclude with a brief mention of applications.

Science, Technology and Society (STS) is a term which can be used very broadly. This review has adhered to the definition provided by Aikenhead (1994, pp 52-53), who defines STS approaches as those which 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; and a philosophical, historical, or social issue within the scientific or technological community.

Scientific literacy is a term frequently encountered in writing about context-based and STS approaches. In the context of this review, it 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 which may affect their lives and the lives of other members of the local, national and global communities of which they are a part.

Ability refers to the capacity to perform specific tasks. Thus, low-ability pupils in this review are equated with those whose performance indicates low levels of achievement.

In making the above distinctions between context-based approaches and STS approaches, it is recognised that there are considerable areas of overlap in terms of the content of materials. However, it is possible to identify some general patterns of use, although with the proviso that there are exceptions. The term ‘context-based’ is more common in Europe, whilst ‘STS’ is more common in North America. ‘Context-based’ is also a term more frequently applied to whole courses, and ‘STS’ to smaller units of enrichment materials. However, there are examples, particularly in the USA of whole courses which are described as STS courses and which share many of the characteristics of courses described as context-based elsewhere.

1.3 Policy and practice background

1.3.1 Interest in context-based and STS approaches

The origins of context-based approaches and their links with STS approaches have been documented by Bennett (2003), and Bennett and Holman (2003). Ideas underpinning the development of context-based approaches have also been documented by Campbell *et al.* (1994). Aikenhead (1994) has produced a detailed overview of STS approaches and materials, and how they draw on contexts and applications to develop ideas about science, technology and society.

The term 'context-based approaches' appears to have been applied to some of the activities in school science classrooms for a little under twenty years. Context-based approaches have their origins in two types of course, those for less academic pupils at the secondary level, such as *Nuffield Science 13-16* (Nuffield Foundation, 1980), and those for non-science specialists at the upper secondary or tertiary level, such as *Science in Society* (Lewis, 1981) in the UK, and *Science and Culture* (Yager and Casteel, 1968) in the USA.

From the early 1980s, context-based approaches started to appear in mainstream science courses and these are now in widespread use in a number of countries. Examples of such materials tend to be found at the secondary and high school level and include, for example, in the UK, *Science: the Salters Approach* (University of York Science Education Group [UYSEG], 1990–1992), *Salters Advanced Chemistry* (Burton *et al.,* 1994) and the SLIP (Supported Learning in Physics Project, 1997). Examples from other countries include *ChemCom* (American Chemical Society [ACS], 1988) in the USA, *PLON* (Dutch Physics Curriculum Development Project, 1988) in the Netherlands, and *STEMS* (Science, Technology Environment in Modern Society) in Israel (Tal *et al.,* 2001).

Courses using context-based and STS approaches are normally characterised by one or both of the following aims: to help young people appreciate how science relates to their current and future lives, and to stimulate interest in science, possibly with a view to encouraging more young people to continue their study of science beyond the compulsory period. Additionally, they tend to be characterised by a broader range of teaching strategies than the traditional teacher exposition and practical work associated with science teaching. Thus context-based and STS approaches advocate the use of small-group discussions, role-plays, poster presentations, problem-solving tasks, creative writing and a variety of activities aimed at encouraging pupils to interact with material they are reading, such as Directed Activities Related to Text (DARTs) (Davies and Green, 1984).

Context-based and STS approaches have attracted interest for a number of reasons, some of which relate to possible benefits they might confer on particular groups of pupils. Early courses – such as, for example, the LAMP (Less Academically-Motivated Pupils) project (ASE, 1978) – were developed specifically for low-ability pupils in order to help them make links between science and their everyday lives. As courses adopting context-based and STS approaches became more widespread in the secondary school science curriculum, they attracted the attention of people working in the area of gender and science, as context-based and STS course embodied many of the recommendations made for making science more attractive to girls. Thus, although not specifically developed with the aim of making science more interesting to girls, one strand of the research into the effects of context-based and STS approaches has focused on gender aspects.

1.3.2 Links with initial teacher training

Initial teacher training courses have tended to promote the use of context-based and STS approaches for several reasons. Firstly, they are seen as motivating for pupils. Secondly, the range of teaching strategies they include gives them strong links with 'pupil-centred learning' or 'active learning'; these terms are applied to activities in which pupils have a significant degree of autonomy over the learning activity, and are frequently advocated in teaching as a means of stimulating pupils' interest in what they are studying (for example, Kyriacou, 1998). Thirdly, and more recently, context-based and STS approaches, with their associated emphasis on discussion and writing tasks, have been promoted as a means for developing pupils' language skills in science, such as in the use of argument (e.g. Newton *et al.,* 1999; Osborne *et al.,* 2001) and the development of writing genres (e.g. Wray and Lewis, 1997).

1.4 Research background

Research into the effects of context-based and STS approaches falls into three main areas. The most significant concerns pupils' affective responses, where one area of focus has been an exploration of effects in relation to gender. For example, a study by Ramsden (1992), gathering data from pupils following a context-based course *(Science: The Salters Approach)*, suggested that pupils who had earlier experience of more conventional courses reported increased enjoyment of science lessons, and pupils following context-based courses cited a wide range of reasons for enjoying science. There was some evidence to support the claim that increased enjoyment of science was reported more frequently by girls than boys. Similar indications emerged from an evaluation of the PLON materials in the Netherlands (Eijkelhof and Kortland, 1988). Interest in gender effects and the use of context-based approaches remains high, as demonstrated by the more recent work undertaken by Sjøberg (2000), who looked in detail at the effects of particular types of context on girls and boys.

A second strand of research has focused on the development of pupils' understanding of scientific ideas as a result of following context-based and STS approaches (e.g. Barker and Millar, 2000). The final strand has explored aspects of teachers' responses to, and use of, materials incorporating context-based and STS approaches (e.g. Borgford, 1995). It is worth noting that, despite the aspirations of context-based and STS courses of making science more attractive to all pupils, comparatively little detailed work has been done exploring effects on lower-ability pupils.

Broadly speaking, two main claims have been made on the basis of research into the effects of context-based and STS approaches:

- They are more effective than conventional approaches at promoting interest in science lessons in pupils generally, with girls and lower-ability learners seeming to benefit in particular.
- They develop an understanding of scientific ideas which is at least as good as that of pupils following more conventional courses.

One purpose of this review is to establish the strength of the evidence base on which these claims are made.

1.5 Authors, funders and other users of the review

1.5.1 Why the Science Review Group is an appropriate group to undertake the review

The members of the Review Group all have considerable research experience in science education, both in the UK and in a number of other countries. Members of the Review Group are also members of the University of York Science Education Group (UYSEG), a group with an international reputation for its research, curriculum development and teacher professional development.

One of the team, Judith Bennett, has recently published a book, *Teaching and Learning Science: a Guide to Recent Research and its Applications* (Bennett, 2003). As Chair of the Research Committee of the Association for Science Education (ASE) from 1999 to 2003, she was directly involved in the promotion of the use of research findings in initial teacher training (ITT) and continuing professional development (CPD) programmes. She also has a significant role in the Science ITT programme in the Department of Educational Studies at York, and was responsible for the introduction at York of a part-time, research-based masters programme, the MA in Science Education, for practising teachers. A number of participants in this programme are teachers associated with the Department of Educational Studies' ITT Partnership, and/or former students on the ITT course. The Review Group has therefore an established track record of working closely with teachers and building research capacity in the area of the review.

This TTA-supported Science Review Group is also part of the EPPI-Centre's DfES-funded Science Review Group, using similar review procedures as required for this review. This Group has recently completed a systematic review of the use of small-group discussions in secondary science teaching (Bennett *et al.*, 2004).

1.5.2 The topicality of the review

The publication of *Beyond 2000* (Millar and Osborne, 1998) stimulated debate over the nature of the school science curriculum and has resulted in moves to increase the emphasis on *scientific literacy*; in other words, to equip pupils with the knowledge, understanding and skills they need as future adult citizens in order to engage with discussion and debate of scientific matters and issues. This overlaps considerably with the aims of context-based and STS approaches. The recent report of the House of Commons Science and Technology Committee (House of Commons, 2002, p 5) highlights the prominent role of scientific literacy: 'A new National Curriculum should require all pupils to be taught the skills of scientific literacy and selected key ideas across the science'. Examples of courses specifically aimed at developing scientific literacy are now emerging, such as *Science for Public Understanding* (Hunt and Millar, 2000), a post-compulsory course for 17- to 18-year-olds, and *21st Century Science*, a GCSE course currently being developed by the University of York and the Nuffield Curriculum Centre.

1.5.3 The potential audiences and users of the findings of the review

The work done for the review, and its findings, are of interest to a number of groups. These include university staff and school staff involved in teaching and supervision on initial teacher training (ITT) courses; the students on such courses; people who write textbooks for such courses and also for school science courses, policy-makers responsible for making decisions about the content of ITT courses, those involved in the development of curriculum materials which emphasise context-based and STS approaches; and parents with an interest in the experiences being offered to their children in school science lessons.

1.6 Review questions

The review has two principal aims:

- to explore the effects of teaching approaches which emphasise placing science in context and promoting links between science, technology and society (STS) on boys and girls, and on lower-ability pupils
- to inform the evidence base on which ITT courses draw in relation to the above teaching approaches

The overall review question is identical to that of the earlier 'base' review:

What evidence is there that teaching approaches which emphasise placing science in context and promote links between science, technology and society improve the understanding of science ideas and the attitudes to science of 11- to 18-year-old pupils, and what are the implications of the evidence for initial teacher training courses?

The focus on understanding encompasses science concepts, ideas about the nature of science, and scientific method. The focus on attitude encompasses attitude towards science, attitude towards school science, motivation to learn, interest in science activities, and career intentions.

From the earlier review, it became apparent that a number of studies made reference to other science-related abilities, such as the development of investigative skills, manipulative skills, communication skills, problem-solving skills and decision-making skills. While this area was not the main focus of the review, such effects were noted when studies were categorised (keyworded) to produce an overview (the systematic map). It is recognised that although some of these skills are not science-specific there is an expectation that they will be applied in science contexts. It is further recognised that there are areas of debate over the interpretation of some of these terms (e.g. attitude, problem-solving skills) and how they might be measured.

In producing the systematic map of the area, it became apparent that the number of studies involved required a narrower focus to be taken for the in-depth review. It was therefore decided to focus the in-depth review on studies of effects on understanding and attitudes of pupils who followed *whole courses* using context-based approaches.

The review questions for the two in-depth review areas are as follows:

What is the evidence from evaluative studies of the effect of context-based or STS courses on the attitude to science and/or the understanding of science ideas of boys and girls in the 11 to 16 age range?

What is the evidence from evaluative studies of the effect of context-based or STS courses on the attitude to science and/or the understanding of science ideas of lower-ability pupils in the 11 to 16 age range?

2. METHODS USED IN THE REVIEW

This review employed the systematic review methods developed by the EPPI-Centre (the Evidence for Policy and Practice Information and Co-ordinating Centre).

There are the following ten main steps in a systematic review of research literature:

- identifying the area for review
- formulating the research question for the review
- identifying criteria for including and excluding studies
- searching for studies which might be suitable for inclusion in the review
- screening the studies to see if they meet the inclusion criteria
- coding each of the included studies (keywording)
- characterising the included studies (systematic mapping)
- selecting an area of the map for in-depth review
- summarising and evaluating the contents of each of the studies included in the in-depth review (data-extraction)
- reporting on the work

The advantages of this strategy are as follows: the process is systematic and thus trustworthy; and it develops research capacity in the Review Group in terms of each part of the process (studies identified, keyworded, and data-extracted).

This review made use of several of the initial stages of the previous 'base' review using an identical general research question. This means that it was not necessary to identify new inclusion/exclusion criteria, search existing databases, screen emerging studies, nor characterise the included studies afresh.

2.1 User-involvement

The timescale for the review precluded wide user-involvement in the identification of the review question and the review process. The Teacher Training Agency (TTA), one of the main users of systematic reviews, was provided with two alternative foci for this review and the current review areas correspond with the TTA's priorities.

2.2 Identifying and describing studies

2.2.1 Defining relevant studies: inclusion and exclusion criteria

No new inclusion and exclusion criteria needed developing. For information, the exclusion criteria used for the previous 'base' review are listed below:

1. Exclusion 1: exclusion on topic (i.e. not science)
2. Exclusion 2: exclusion on aspects of science (not science in context or STS)
3. Exclusion 3: exclusion on learning effects (not understanding or attitude)

4. Exclusion 4: age (not 11–18)
5. Exclusion 5: exclusion on study type:
   (a) editorials, commentories, book reviews
   (b) policy documents
   (c) resources
   (d) bibliography
   (e) theoretical (non-empirical) papers
   (f) methodology papers
   (g) review papers
6. Exclusion 6: language (not English)
7. Exclusion 7: date (not 1980–2002)

Appendix 2.1 provides the sheet used when applying the inclusion and exclusion criteria.

2.2.2 Identification of potential studies: search strategy

The search strategy for identifying potential studies used in the previous 'base' review has been maintained, as described below:

- a hierarchy of bibliographic databases in the following order: ERIC, BEI, PsycInfo and SSCI (see Appendix 2.2 for details of the search strategy used)
- handsearching of the journal *Studies in Science Education*, which specialises in major review articles in science education, in order to identify in bibliographies any potentially useful studies
- personal contacts
- direct requests to key informants, following completion of electronic and handsearching

2.2.3 Screening studies: applying inclusion and exclusion criteria

The screening strategy used in the previous 'base' review has been maintained. The Review Group set up a database system (using *EndNote* software) for keeping track of, and coding, studies found during the review. Inclusion and exclusion criteria were applied to titles and abstracts. Studies excluded on the basis of titles and abstracts were recorded on the database, together with reasons for their exclusion. Inclusion and exclusion criteria were re-applied to the full reports and those which did not meet the initial criteria were excluded.

Due to limited resources, no attempts were made to obtain copies of dissertations from abroad.

The process of searching and screening yielded 67 papers reporting a total of 61 studies. Although this review is interested in documenting and evaluating research studies, the initial review stages of searching and screening will focus on review papers. Several papers may report on the same study. For the purposes of this review, we consider papers to report on the same study if the papers use identical samples and data-collection methods, and analyse the same, or a subset of the same, data. The use of a similar data-collection method (with or without the same analysis method) with a subsequent cohort of learners would constitute a new study.

Whereas the map in the 'base' review (Bennett *et al*., 2003) was presented for individual papers, the map in this review will be presented as an overview of characteristics of research studies, where applicable, based on keywords of combination of papers reporting the same study. Equally, the in-depth review makes use of evidence from research studies based on data-extraction.

2.2.4 Characterising included studies

The keywording strategy used for the previous 'base' review was maintained. The purpose of this process is to describe the characteristics of studies, as opposed to summarising and judging the quality of the findings.

Keywording involves applying standard keywords supported by the EPPI-Centre keywording strategy (EPPI-Centre, 2002a). These keywords are given in Appendix 2.3 and are used to keyword all systematic reviews which use EPPI-Centre methods.

It needs to be noted that this review used a particular interpretation of evaluations, one of the study types included in category 10 of the core keywords (see Appendix 2.3). The EPPI-Centre guidelines consider two sorts of evaluation (1) where the intervention being evaluated occurs as a normal part of practice which may then be evaluated *post hoc* ('naturally occurring') and (2) where the intervention is introduced specifically in order to be evaluated, that is, solely as part of research ('researcher-manipulated'). Both types of evaluations may or may not have control groups outside the intervention. For this review, however, evaluations where a researcher actively allocates pupils to different learning approaches for the purposes of the research seemed unrealistic for large-scale curriculum interventions, but the issue of whether or not there was a control group was considered particularly relevant. Hence, 'evaluations of researcher-manipulated interventions' was redefined for the purposes of this review to mean studies which focus on pupils taking existing programmes with a different learning approach, and with a control group as part of the research design. In contrast, the term 'evaluations of naturally occurring interventions' referred to evaluations in which there was no control group.

The EPPI-Centre core keywords were supplemented by review-specific keywords, developed to characterise aspects of studies central to the main review question. These keywords are given in Appendix 2.3.

The data gathered through keywording were recorded on the electronic database. Once studies had been keyworded, they were scrutinised using the electronic database in order to produce the systematic map. The systematic map, based on 61 studies, differs in important details from the map included in the previous 'base' review. The latter summarised 66 papers, of which five pairs reported on the same study.

In the map the following areas were scrutinised:

- linked reports
- the country of study
- the age/level of the pupils
- the type of study
- the discipline of the study
- the nature of the intervention
- the aspects of science learning the study was assessing
- the outcome measures
- the outcomes

The systematic map is presented in sections 3.1 and 3.2. The numbers of studies involved at various stages of this review are presented in Figure 3.1.

2.2.5 Identifying and describing studies: quality-assurance process

During the previous 'base' review, various quality-assurance procedures were used. The application of inclusion and exclusion criteria was initially conducted by three team members for a 2.5% random sample (58 studies). This was done independently in the first instance, and the team members then met to compare the codes allocated, discuss the discrepancies, and reach a consensus on how criteria were to be interpreted and applied. Two team members and a member of the EPPI-Centre then worked on a second 2.5% random sample (58 studies), working independently. These data were used to calculate inter-screener agreement, using frequency counts and the Cohen's Kappa inter-screener reliability coefficients.

Once the 61 studies to be included in the review had been identified, a 10% sample (six studies) was keyworded by three team members to check the appropriateness of the review-specific keywords and reach a consensus on how keywords were to be applied. Following minor modifications to the review-specific keywords, thus increasing the validity, a 29% sample (16 studies) was double-keyworded in order to establish reliability. A member of the EPPI-Centre also keyworded studies for a random sample of nine studies.

2.3 In-depth review

2.3.1 Moving from broad characterisation (mapping) to in-depth review

The purpose of in-depth reviewing is to describe the characteristics of studies in more detail, and to assess the quality of methods used and the findings of studies. An in-depth review involves summarising and evaluating the contents of each of the studies included in the in-depth review (data-extraction).

The inclusion criteria for studies in the in-depth review are as follows:

- The studies are evaluations of whole courses which used context-based or STS approaches.

- The studies evaluate interventions for pupils in the 11 to 16 age range.

- The studies report outcomes in relation to gender and/or lower ability of pupils.

For the purposes of this review, a course was defined as a coherent learning programme with an explicit curriculum outline which is subject to an accredited examining system. In contrast, units or modules of learning which are inserted into other programmes, as identified through the review-specific keywords in Appendix 2.3, were excluded.

Thus studies were excluded from the in-depth review on the following bases:

1. Exclusion 1: exclusion on focus (i.e. study is not about whole courses using context-based or STS approaches)

2. Exclusion 2: exclusion on age range (i.e. study is not about interventions for 11- to 16-year-olds)

3. Exclusion 3: exclusion on outcomes (i.e. study is not reporting findings on the effect on boys and girls, and/or on lower-ability pupils)

2.3.2 Detailed description of studies in the in-depth review

Studies identified as meeting the inclusion criteria for in-depth review were double data-extracted and quality assessed, using the EPPI-Centre's detailed data-extraction guidelines (EPPI-Centre, 2002c) and software, EPPI-Reviewer (EPPI-Centre, 2002b).

2.3.3 Assessing quality of studies and weight of evidence for the review question

Once data have been extracted from the studies, the next step in the review is to assess the quality of the studies and the weight of evidence they present in relation to the review question. The EPPI-Centre data extraction procedures identify three initial categories (A, B, C) and a fourth overall category (D) referred to as weight of evidence (WoE) to help in the process of apportioning different weights to the findings and conclusions of different studies, with each of these being rated as high, medium or low.

The categories are as follows:

Category A: Trustworthiness of findings (internal methodological coherence) in relation to the study question

Category B: Appropriateness of the research design and analysis used for answering the review question

Category C: Relevance of the study topic focus (from the sample, measures, scenario, or other indicator of the focus of the study) to the review question

Finally, an overall weighting (category D) is compiled based on the judgements reached in categories A, B and C above.

For category A, the M.11 judgement from the EPPI-Centre data-extraction guidelines (EPPI-Centre 2002c) was used, ranging from high, through medium high, medium and medium low, to low.

Judgements of weighting in category B and C are based on the quality of the study's research work related to the appropriateness for this specific in-depth review. Separate but partly overlapping sets of weight of evidence (WoE) criteria were developed for each of the four areas of this review: that is, gender affecting attitude (GA), gender affecting understanding (GU), lower ability affecting attitude (LA), and lower ability affecting understanding (LU). Appendix 2.4 shows the WoEs for each of the four areas of this review. They indicate how the Review Group interpreted the appropriateness of the research design and analysis (category B) through five aspects: that is, the sample size/sampling method; nature of a comparison group for gender/lower ability; benchmark data on attitude/understanding; the reliability/validity of the data-collection method; and the reliability/validity of the data-analysis method. Each of these aspects has three level descriptors with a weighting of 3, 2 or 1 in decreasing appropriateness. The sum total of the weighted aspects determines the overall weight of category B as follows:

5–6 = low
7–8 = medium-low
9–11 = medium
12–13 = medium-high
14–15 = high

Similarly, Appendix 2.4 shows how the relevance of the study topic focus (category C) has been weighted through five aspects: that is, the representativeness of the sample for gender/lower ability; the focus of the intervention on gender/lower ability; the appropriateness of the measures for attitude/understanding; the breadth of testing attitude/understanding; the representativeness of the classroom situation. Again, each of these aspects has three level descriptors with a weighting of 3, 2 or 1 in decreasing appropriateness. The sum total of the weighted aspects determines the overall weight of category C in the same way as explained for category B above.

The total weighting for category D was constructed by the Review Group by allocating equal weighting to judgements made for A, B and C, each ranging from 1 (low) to 5 (high). Sum totals for D were classified as follows:

Total sum score	D classification
3–4	low
5–7	medium low
8–10	medium
11–13	medium high
14–15	high

2.3.4 Synthesis of evidence

The synthesis consolidating evidence from this review draws primarily on the findings from studies weighted as *high*, *medium-high* and *medium*. Findings from studies weighted as *medium-low* are only considered if these corroborate findings of studies with a higher weight of evidence.

2.3.5 In-depth review: quality-assurance process

Once the 14 studies to be included in the in-depth review had been identified, a further check was undertaken by two team members on these studies to ensure that keywording had been done consistently and accurately. Other studies, which appeared to come close to the inclusion criteria, were also double-checked to ensure that appropriate decisions had been made.

A summary table (Appendix 4.1) was used for initial data-extraction. Three team members undertook a data-extraction on one of the in-depth review studies, working first individually and then meeting to moderate their summaries. This process increased the reliability of the subsequent data-extractions. It also served to adjust the summary table format and refine the WoE criteria sheets (see Appendix 2.4). Data-extraction was then conducted by pairs of team members, working first independently and then comparing their decisions and coming to a consensus. Such consolidated data-extractions were uploaded on EPPI-Reviewer. In addition, for purposes of quality assurance, two members of the EPPI-Centre have also previously double data-extracted and quality assessed three of the studies included in the in-depth review and part of the previous ‘base’ review (Bennett *et al.*, 2003).

Members of the Review Group discussed the consistency of the application of criteria for judgements of quality across the in-depth review as a whole. The validity of the WoE criteria sheets was checked by each team member independently assigning a ranking to the 14 studies, and these were then compared.

3. IDENTIFYING AND DESCRIBING STUDIES: RESULTS

Sixty-one studies were included in the review after searching and screening had taken place. Figure 3.1 shows how these studies were identified from searching to map to synthesis.

Figure 3.2 presents the origin of each of the 61 studies included in the map. The hierarchy of electronic databases described in section 2.2.2 was used in compiling the electronic database entries in this table: that is, ERIC entries were imported first, followed by BEI entries, then entries from PsycInfo, and finally entries from SSCI. Any duplicated entries were removed from the electronic database of review papers. Thus, if a study appeared as an entry in ERIC, and again in BEI, its source is given as ERIC.

Figure 3.2: The sources of the 61 studies included in the systematic map

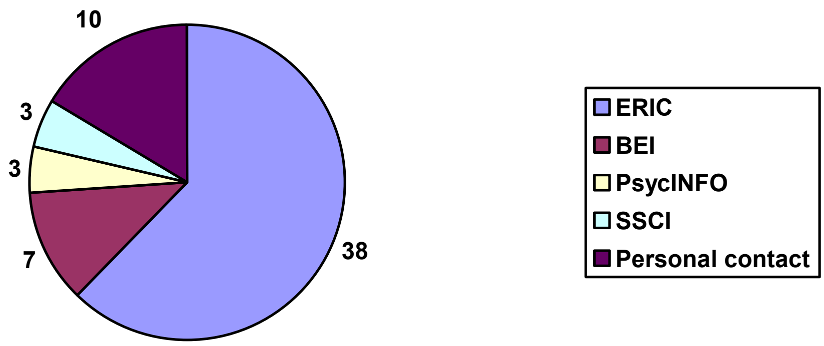
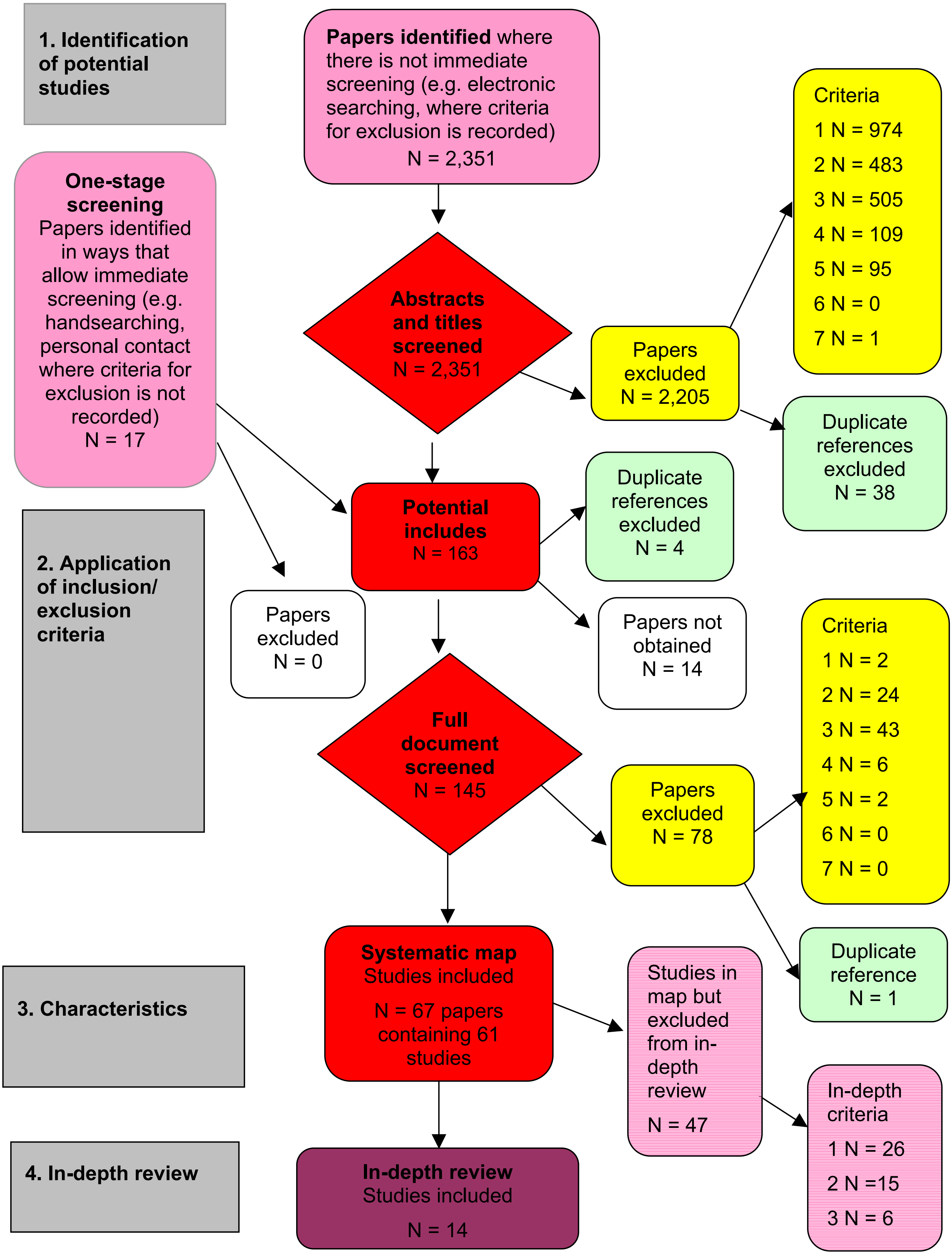


Figure 3.1: Filtering of papers from searching to map to synthesis



3.1 Studies included from searching and screening

As described in section 2.2.4, the following areas were scrutinised in detail to produce the systematic map:

- linked reports
- the country of study
- the age/level of the pupils
- the type of study
- the discipline of the study
- the nature of the intervention
- the aspects of science learning the study was assessing
- the outcome measures
- the outcomes

Of the 61 studies, 57 were published studies, with the remainder being two unpublished MA theses, one unpublished PhD thesis, and one set of conference proceedings. The nature of the review question meant that there was uniformity in responses to a number of the keywording categories. All the studies focused on curriculum, and all also focused on pupils. Within this group, 12 focused on teachers as well as pupils. The majority of the studies were undertaken in mixed-sex educational settings, with four in single-sex settings, all of which were girls' schools.

3.2 Characteristics of the included studies

3.2.1 Related research activity

There were a number of productive groups of research in the review 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 co-ordinated research programmes. In particular, there were clusters of studies undertaken by:

- researchers at the University of Utrecht in the Netherlands on the Dutch Physics Curriculum Project (PLON) (three papers)

- researchers at the University of York Science Education Group on the Salters curriculum development projects (six papers) and context-based materials developed for use in southern Africa (three papers)

- Yager and co-workers in Iowa the USA on a variety of aspects of the use of STS materials (two studies)

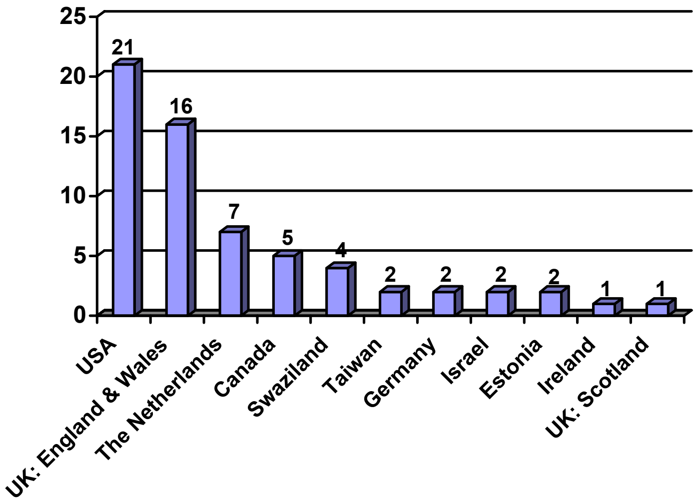
- Zoller and co-workers in Israel and British Columbia in Canada on STS courses for upper high-school students (three studies)

- Solomon in the UK on the discussion of socio-scientific issues (three papers)

3.2.2 The country of study

Figure 3.3 shows the countries in which the studies were undertaken. The total number of countries is 63, as two studies were carried out in more than one country. Over 80 percent were carried out in the USA, the UK, the Netherlands and Canada.

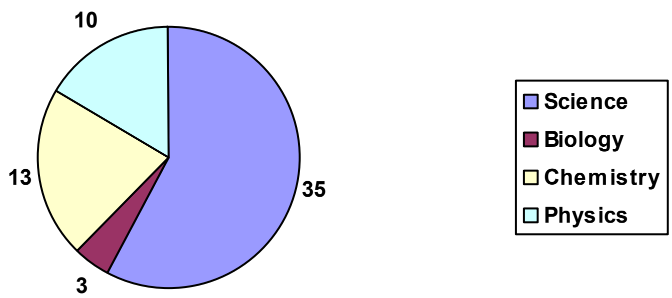
Figure 3.3: The country of study (N = 61, not mutually exclusive, total = 63)



3.2.3 The discipline of the study

Figure 3.4 shows the science discipline of the study.

Figure 3.4: The science discipline of study (N = 61, mutually exclusive)



It is likely that the focus on chemistry and physics in the individual science disciplines in part 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.

3.2.4 The age/level of the pupils

Forty-one studies were undertaken with pupils in the 11–16 age range, and 18 with pupils in the 17–20 age range. Within this, three studies had also gathered data from the 6–10 age range, and one from the age range 21 and over. Two studies did not give details of the age range of the pupils. The emphasis on pupils 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.

3.2.5 The type of study

Twenty-four of the studies were evaluations of researcher-manipulated policies or practices. Thirty-seven of the studies were evaluations of naturally occurring policy and practices.

3.2.6 The nature of the intervention

Figure 3.5 shows the nature of the intervention described in each of the studies.

Figure 3.5: The nature of the intervention (N = 61, mutually exclusive)

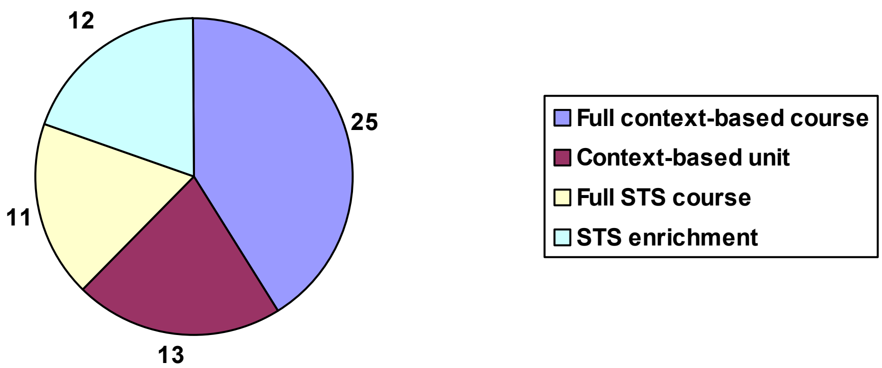


Figure 3.5 shows that 62% (38 out of 61) of the studies evaluated context-based interventions, with two out of every three of these dealing with full context-based courses. The remaining 38% of the studies report on evaluations of STS approaches. In contrast with context-based interventions, the majority of evaluations of STS approaches focus on STS enrichment activities within traditional approaches.

Table 3.1: The nature of the intervention in relation to the age range (N = 61, mutually exclusive)

	Total	11–16 age group (N = 41)	17–18 age group (N = 18)	Age group not stated (N = 2)
Full context-based course	25	14	11	-
Context-based unit	13	12	-	1
Full STS course	11	7	4	-
STS enrichment	12	8	3	1
Total	***61***	***41***	***18***	***2***

Table 3.1 indicates that studies evaluating the effect of full courses are over-represented amongst those dealing with curriculum for the 17–18 age group. This applies to evaluations of full context-based and full STS courses, and may be explained by the interest in the relationship of the teaching approach with the curriculum final assessment. In contrast, studies on smaller scale interventions, such as the introduction of context-based units and STS enrichment, are over-represented amongst the 11–16 age group. The outcomes in these cases may be intended to influence classroom teaching more directly.

Table 3.2: The nature of the intervention in relation to the nature of evaluation (N = 61, mutually exclusive)

	Total	Researcher manipulated (N = 24)	Naturally occurring (N = 37)
Full context-based course	25	11	14
Context-based unit	13	2	11
Full STS course	11	7	4
STS enrichment	12	4	8
Total	***61***	***24***	***37***

Table 3.2 suggests that evaluations of full STS courses are relatively more frequently researcher-manipulated in nature, whereas evaluations of the two types of small-scale interventions use relatively more frequently a naturally-occurring design. Studies of full context-based courses do not have a preference for any of the two evaluation designs.

3.2.7 The outcome measures

Figure 3.6 shows the outcome measures used in the studies. Figures total more than 61 as several studies employed more than one outcome measure.

Figure 3.6: Outcome measures (N = 61, not mutually exclusive, total = 116)

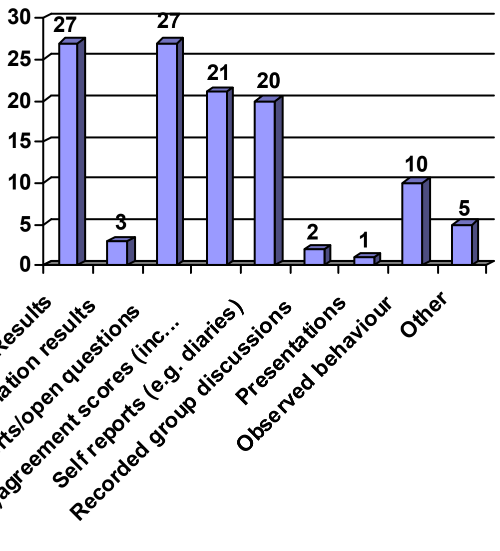


Figure 3.6 shows that just over two in five of the studies used tests or written reports. One in three of the studies used (dis)agreement scores, including Views on Science-Technology-Society (VOSTS) or self-reports.

Table 3.3: Outcome measures in relation to nature of evaluation (N = 61, not mutually exclusive)

	Total	Researcher manipulated (N = 24)	Naturally occurring (N = 37)
Test results	27	15	12
External examination results	3	-	3
Written reports/open questionnaires	27	11	16
Concept webs	-	-	-
(Dis)agreement scores (including VOSTS)	21	10	11
Self reports (e.g. diaries, interviews)	20	4	16
Recorded group discussions	2	1	1
Presentations	1	-	1
Observed behaviour	10	3	7
Other	5	4	1
Totals	***116***	***48***	***68***

Table 3.3 shows the outcome measures as used in researcher-manipulated and naturally occurring evaluations. Of the four most common measures, only self-reports (most usually interviews) emerged as being used much more frequently in naturally-occurring evaluations than in researcher-manipulated evaluations; 16 out of 37 (around 43%), compared with 4 out of 24 (around 17%). The reverse was the case for the use of test results, where almost two-thirds (15 out of 24) of the researcher-manipulated interventions used this measure, compared with one-third (12 out of 37) of the naturally-occurring interventions.

3.2.8 The outcomes

Of the studies, 44 reported on attitudes and 41 on understanding. Of these, 24 reported on both these aspects. Two other aspects which also emerged as featuring prominently in studies were the development of skills (21 studies), and 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, where science is mostly taken as a compulsory subject.

Having looked at individual categories on the map, it was decided to undertake a cross-tabulation of selected categories to help with the process of identifying the studies for the in-depth review. The cross-tabulation focused on the following characteristics:

- the nature of the intervention
- the age range
- the science focus
- the name of the course (to establish if any courses were receiving particular attention in research studies)
- the nature of the outcomes reported

In relation to this last characteristic, in addition to noting outcomes in relation to attitude and understanding, it was decided to note outcomes in relation to lower ability and to gender, as these had emerged as aspects featuring prominently in a number of studies.

Tables 3.4 and 3.5 summarise these data. A comparison of these tables shows that studies on context-based/STS approaches in (integrated) science or the different separate disciplines are equally spread over the researcher-manipulated and naturally-occurring evaluations. Within each of these clusters, two in three studies focus on a science curriculum (15 out of 24, and 23 out of 37 studies, respectively).

Similarly, studies focusing on the two age-ranges are also more or less equally distributed over the two types of evaluations. Studies on the 11–16 age range constitute just over two out of three of the researcher-manipulated and naturally-occurring evaluations (i.e. 17 out of 24, and 24 out of 37 studies, respectively). The remaining third of each type of evaluation focuses on the 17–18 age range.

Table 3.4: Researcher-manipulated evaluations: outcomes measured
Total number of studies = 24 (11 to 16 age group N = 17; 17 to 18 age group N = 7)

Nature of interaction	Age range	Paper	Outcomes measured: Attitude	Outcomes measured: Under-standing	Outcomes measured: Lower ability	Outcomes measured: Gender	Science	Course
Full context (N = 11)	11–16 (N = 6)	Gallagher *et al.* (1992)	✓				Science	
		Ramsden (1997)	✓	✓		✓	Science	Science: The Salters Approach
		Smith and Bitner (1993)		✓		✓	Chemistry	ChemCom
		Wierstra (1984)	✓	✓		✓	Physics	PLON
		Wiestra and Wubbels (1994)	✓	✓		✓	Physics	PLON
		Yager and Weld (1999)	✓	✓		✓	Science	Scope, Sequence and Continuity
	17–18 (N = 5)	Banks (1997)		✓			Chemistry	Salters Advanced Chemistry
		Barber (2000)	✓	✓			Chemistry	Salters Advanced Chemistry
		Barker and Millar (1996)		✓			Chemistry	Salters Advanced Chemistry
		Key (1998)	✓				Chemistry	Salters Advanced Chemistry
		Winther and Volk (1994)		✓			Chemistry	Salters Advanced Chemistry
Context-based unit	11–16 (N = 2)	Dahncke *et al.* (2001)		✓			Physics	
		Lubben *et al.* (1997)		✓		✓	Science	Matsapha materials

Table 3.4: Researcher-manipulated evaluations: outcomes measured (continued)
Total number of studies = 26 (11 to 16 age group N = 18; 17 to 18 age group N = 8)

Nature of interaction	Age range	Paper	Outcomes				Science	Course
			Attitude	Under-standing	Lower ability	Gender		
Full STS courses (N = 7)	11–16 (N = 5)	Ben-Zvi (1999)	✓	✓			Science	Science and Technology for all
		Smith and Matthews (2000)	✓		✓	✓	Science	
		Tsai (2000)		✓			Science	
		Yager *et al.* (1988)	✓	✓			Science	
		Zoller *et al.* (1990)	✓	✓		✓	Science	ST 11
	17–18 (N = 2)	Solomon (1992)	✓				Science	Discussions of Issues in School Science (DISS)
		Zoller *et al.* (1991)	✓				Science	ST 11
STS enrichment (N = 4)	11–16 (N = 4)	Christie and Nelson (1988)	✓				Science	
		Rubba (1990)	✓				Science	
		Rubba *et al.* (1991)	✓	✓			Science	Genetics unit
		Wiesenmeyer and Rubba (1999)	✓				Sciences	

Table 3.5: Naturally-occurring evaluations: outcomes measured
Total number of studies = 37 (11 to 16 age group N = 24; 17 to 18 age group N = 11; age group not stated N = 2)

Nature of intervention	Age range	Paper	Outcomes reported				Science	Course
			Attitude	Under-standing	Lower ability	Gender		
Full context N = 14	11–16 (N = 8)	Eijkelhof and Lijnse (1988)	✓	✓			Physics	PLON
		Greenlee and Lambert (1996)	✓			✓	Science	The Science of Living Spaces
		Huppert *et al.* (1992)	✓	✓	✓		Biology	Human Health and Science
		Licht (1990)	✓				Physics	PLON
		Nentwig *et al.* (2002)	✓		✓	✓	Chemistry	Chemie im Kontext
		Ramsden (1992)	✓	✓		✓	Science	Science: The Salters Approach
		Reid and Skryabina (2002)	✓			✓	Physics	Standard Grade Applications-led course
		Sutman and Bruce (1992)	✓	✓	✓		Chemistry	ChemCom
	17–18 (N = 6)	Barker and Millar (2000)		✓			Chemistry	Salters Advanced Chemistry
		Campbell *et al.* (2000a)	✓	✓			Physics	Salters Horners Advanced Physics
		Hughes (2000a)	✓	✓		✓	Chemistry	Salters Advanced Chemistry
		Pedersen (1992)	✓	✓			Chemistry	ChemCom
		Pilling *et al.* (2001)	✓	✓		✓	Chemistry	Salters Advanced Chemistry
		Truex (1987)	✓	✓	✓		Science	Contemporary Science

Table 3.5: Naturally-occurring evaluations: outcomes measured (continued)
Total number of studies = 37 (11 to 16 age group N = 24; 17 to 18 age group N = 11; age group not stated N = 2)

Nature of intervention	Age range	Paper	Outcomes reported: Attitude	Under-Standing	Lower ability	Gender	Science	Course
Context-based unit (N = 11)	11–16 (N = 10)	Campbell *et al.* (2000b)		✓				Linking School Science with Industry and Technology
		Dlamini *et al.* (1996)	✓			✓	Science	Matsapha
		Helms (1998)	✓				Science	Wetlands Restoration
		Kortland (1997)		✓			Physics	Garbage reusing/recycling
		Kortland (2001)		✓			Science	
		Lubben *et al.* (1996)	✓	✓			Science	Matsapha
		Monhardt and Monhardt (1998)		✓			Science	Jurisprudential Inquiry model to STS
		Rannikmäe (2001)	✓	✓			Science	
		Roth (2001)		✓	✓		Science	
		Van Weelie (2001)		✓			Science	Biodiversity
	No age (N = 1)	Fowler and Sinclair (1992)	✓	✓			Science	

Table 3.5: Naturally-occurring evaluations: outcomes measured (continued)
Total number of studies = 37 (11 to 16 age group N = 24; 17 to 18 age group N = 11; age group not stated N = 2)

Nature of intervention	Age range	Paper	Outcomes reported: Attitude	Under-Standing	Lower ability	Gender	Science	Course
Full STS courses (N = 4)	11–16 (N = 2)	Ebenezer and Zoller (1993)	✓			✓	STS	ST 11
		Wiley (1991)	✓				STS	
	17–18 (N = 2)	Solomon (1988)		✓			STS	
		Solomon (1989)	✓				STS	Discussions of Issues in School Science (DISS)
STS enrichment (N = 8)	11–16 (N = 4)	Ferreira (2001)	✓				Physics	
		Stevens (2001)	✓				Science	
		Walker (1990)	✓				Science	SATIS
		Yager and Tamir (1993)	✓	✓	✓	✓	Science	Iowa STS Model
	17–18 (N = 3)	Alsop *et al.* (1998)		✓			Physics	
		Lenton (1991)	✓	✓			Science	SATIS
		Phillips and Norris (1999)		✓			Science	
	No age (N = 1)	Williams *et al.* (1990)	✓	✓			Science	The Illinois River Project

3.3 Identifying and describing studies: quality-assurance results

The quality-assurance procedures for screening and keywording described in section 2.2.5 were followed as part of the previous ‘base’ review and fully reported there. In brief, inter-screener agreement for a random 2.5% sample (58 studies) for screening studies on the basis of title and abstract varied from 83% to 91%, with fair to moderate Cohen’s Kappa coefficients. Inter-rater agreement of keywording of a 20% sample (16 studies) varied between 79% and 97%, apart from the agreement between non-science and science specialists yielding an inter-rater agreement of just less than 50% for review specific keywords. This low agreement seems due to the difference in familiarity with science education learning outcomes, curriculum initiatives and the possible variations in context-based teaching.

4. IN-DEPTH REVIEW: RESULTS

4.1 Selecting the studies for the in-depth review

The systematic map, which was based on 61 studies, allowed the identification of studies which report on the effect of context-based/STS approaches in science teaching on gender and lower-ability pupils. Applying the exclusion criteria described in section 2.3.1 resulted in 14 studies for the in-depth review.

Two studies (Hughes, 2000; Pilling *et al.*, 2001) have been excluded from the in-depth review because they report on the effect of full context-based courses on boys and girls, but for older pupils. Conversely, four studies (Dlamini *et al.*, 1996; Lubben *et al.*, 1997; Roth, 2001; Yager and Tamir, 1993) report on gender or lower-ability aspects for the target age range but are excluded since they evaluate context-based units rather than full courses.

4.2 Comparing the studies selected for in-depth review with the studies in the systematic map

Studies yielded

The 14 studies included in the in-depth review are as follows:

Ebenezer JV, Zoller U (1993) Grade 10 students' perceptions of and attitudes toward science teaching and school science. *Journal of Research in Science Teaching* **30**: 175–186.
Greenlee S, Lambert L (1996) *The Science of Living Spaces: Women in the Environment of the 21st Century. An Evaluation and Guide Book.* Virginia, USA: Christopher Newport University.
Huppert J, Simchoni D, Lazarowitz R (1992) Human health and science: a model for an STS high school biology course. *American Biology Teacher* **54**: 395–400.
Nentwig P, Parchmann I, Demuth R, Graesel C, Ralle B (2002) Chemie im Kontext: from situated learning in relevant contexts to a systematic development of basic chemical concepts. Paper presented at the Symposium on Context-Based Science Curricula. Kiel, Germany: October 10–13.
Ramsden JM (1992) If it's enjoyable, is it science? *School Science Review* **73**: 65–71.
Ramsden JM (1997) How does a context-based approach influence understanding of key chemical ideas at 16+? *International Journal of Science Education* **19**: 697–710.
Reid N, Skryabina E (2002) Attitudes towards physics. *Research in Science and Technological Education* **20**: 67–81.
Smith G, Matthews P (2000) Science, technology and society in transition year: a pilot study. *Irish Educational Studies* **19**: 107–119.
Smith LA, Bitner BL (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. Kansas City, MO, USA: April 1–4.

Sutman F, Bruce M (1992) Chemistry in the community – Chemcom. *Journal of Chemical Education* **69**: 564–567.
Wierstra RFA (1984) A study on classroom environment and on cognitive and affective outcomes of the PLON-curriculum. *Studies in Educational Evaluation* **10**: 273–282.
Wierstra RFA, 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**: 437–455.
Yager RE, Weld JD (1999) Scope, sequence and co-ordination: The Iowa Project, a national reform effort in the USA. *International Journal of Science Education* **21**: 169–194.
Zoller U, Ebenezer JV, Morley K, Paras S, Sandberg V, West C, Wolthers T, Tan SH (1990) Goal attainment in science-technology-society (S/T/S) education and reality: the case of British Columbia. *Science Education* **74**: 19–36.

Two of these studies were reported in linked pairs of papers. One paper was selected as the lead paper for each study but data in both papers were drawn on for data-extraction purposes. The linked pairs of papers are as follows:

- Reid and Skryabina (2002) and *Skryabina (2000)
- Zoller *et al.* (1990) and *Zoller *et al.* (1991)

Full references for subsidiary papers (asterisked*) are given in the bibliography in Section 6 of the report. For the remainder of this section of the report and throughout the findings and conclusions in Chapter 5, the lead paper only is cited.

Studies' focus of interest

Nine of the 14 studies report on aspects of gender, two on lower ability and three on both of these. Table 4.1 provides an overview of the studies, including gender aspects exploring impact on pupils' attitude (GA) or understanding (GU), and the studies reporting on aspects of lower ability and the impact on pupils' attitude (LA) or understanding (LU).

Table 4.1: Overview of evaluation aspects included in the in-depth review per study

Authors, year	Evaluation aspect: Gender – Attitude (GA)	Gender – Understanding (GU)	Lower ability – Attitude (LA)	Lower ability – Understanding (LU)
Ebenezer and Zoller (1993)	✓			
Greenlee and Lambert (1996)	✓			
Huppert *et al.* (1992)			✓	✓
Nentwig *et al.* (2002)	✓		✓	
Ramsden (1992)	✓			
Ramsden (1997)	✓			
Reid and Skryabina (2002)	✓			
Smith and Bitner (1993)		✓		
Smith and Matthews (2000)	✓		✓	
Sutman and Bruce (1992)				✓
Wierstra (1984)	✓	✓		
Wierstra and Wubbels (1994)	✓	✓		
Yager and Weld (1999)	✓	✓	✓	✓
Zoller *et al.* (1990)	✓			
Total frequencies	***11***	***4***	***4***	***3***

Countries of studies

Table 4.2 shows the countries in which studies selected for the in-depth review were carried out. Almost one-third of the studies were undertaken in the USA, with others as detailed below. The proportions reflect the national distribution in the systematic map. It is reassuring that evaluations of all major context-based/STS curriculum initiatives known to the Review Group are represented in the in-depth review: that is, the Iowa-based SS&C curriculum and ChemCom in the USA, Salter Science curricula in England and Wales, the British Columbia Secondary Science programme in Canada, and the PLON courses in the Netherlands.

Table 4.2: Countries in which the studies selected for in-depth review were carried out (N = 14, mutually exclusive)

Country	Number of studies	Study
USA	4	Greenlee and Lambert (1996) Smith and Bitner (1993) Sutman and Bruce (1992) Yager and Weld (1999)
England and Wales	2	Ramsden (1992) Ramsden (1997)
Canada	2	Ebenezer and Zoller (1993) Zoller *et al.* (1990)
Netherlands	2	Wierstra (1984) Wierstra and Wubbels (1994)
Germany	1	Nentwig *et al.* (2002)
Ireland	1	Smith and Metthews (2000)
Israel	1	Huppert *et al.* (1992)
Scotland	1	Reid and Skryabina (2002)

The researchers

Most of the 14 studies appeared to be undertaken by researchers based in university departments closely involved in the development of the evaluated curriculum or in the introduction of teachers to the context-based/STS curriculum. Two of the studies involved a school teacher as co-author (Huppert *et al.*, 1992; Smith and Bitner, 1993). Three other studies seemingly resulted from higher degree studies (Reid and Skryabina, 2002; Smith and Matthews, 2000; Wierstra and Wubbels, 1994).

Subject focus

Six of the 14 studies in the in-depth review focused on context-based/STS approaches of teaching science (Ebenezer and Zoller, 1993; Greenlee and Lambert, 1996; Ramsden, 1992; Smith and Matthews, 2000; Yager and Weld, 1999; Zoller *et al.*, 1990). Four studies report on a context-based/STS approach in teaching chemistry (Nentwig *et al.*, 2002; Ramsden, 1997; Smith and Bitner, 1993; Sutman and Bruce, 1992), three in teaching physics (Reid and Skryabina, 2002; Wierstra, 1984; Wierstra and Wubbels, 1994) and one in teaching biology (Huppert *et al.*, 1992). This constitutes a lower proportion of science courses and slightly higher proportions of courses in chemistry and physics in comparison with the studies in the map. This difference is probably due to concern about low uptake by girls of chemistry and physics courses in particular, generating a larger research activity in evaluating gender aspects for these disciplines.

Evaluations of courses using an STS approach are under-represented in the in-depth review with only three studies (Ebenezer and Zoller, 1993; Smith and Matthews, 2000; Zoller *et al.*, 1990), while the remaining 11 studies report on context-based courses.

Type of study

Half the studies were naturally occurring evaluations (Ebenezer and Zoller, 1993; Greenlee and Lambert, 1996; Huppert *et al.*, 1992; Nentwig *et al.*, 2002;

Ramsden 1992; Reid and Skryabina, 2002; Sutman and Bruce, 1992); the other half consist of researcher-manipulated evaluations of policies and practices (Ramsden, 1997; Smith and Bitner, 1993; Smith and Matthews, 2000; Wierstra, 1984; Wierstra and Wubbles, 1994; Yager and Weld, 1999; Zoller *et al.*, 1990). All the studies in the latter group were controlled trials.

4.3 Further details of studies included in the in-depth review

4.3.1 Overview of the studies

This section gives an overview of the data extracted from the 14 studies. Section 4.3.1 provides an overview of the aims of the studies, and, in section 4.3.2, methodological considerations are synthesised in order to permit judgements to be reached about the quality of the studies (weight of evidence A).

Section 4.3.3 looks at the research design of the studies in relation to the in-depth review question in order to permit judgements to be reached about the appropriateness of the study design for the in-depth review question (weight of evidence B).

Section 4.3.4 addresses the relevance of the focus of the studies for the in-depth review question in order to permit judgements to be reached about the relevance to the in-depth review question (weight of evidence C).

Aims of studies

All but one of the studies focused on evaluating specific intervention programmes using a context-based or STS approach. The exception, the study by Reid and Skryabina (2002), aimed primarily at documenting attitudes to school science and physics courses (and changes therein) of pupils in Scotland aged 10 to 20 years, and in particular, identifying factors making physics at Higher Grade a popular subject choice. The fact that the Scottish pre-Higher Grade curriculum uses a context-based approach made this study highly relevant for this review.

Five studies intended to evaluate specific learning outcomes of courses using a context-based or STS approach:

- Greenlee and Lambert (1996): To what extent does the intervention encourage girls to consider careers in science, engineering or mathematics? What aspects of the intervention are considered to be encouraging girls to consider such a career?
- Smith and Bitner (1993): How does learning through ChemCom affect the use of formal operational thinking and students' reasoning levels? How does reasoning ability differ between boys and girls?
- Smith and Matthews (2000): How effective is the intervention in improving students' attitudes to science and technology, and in increasing students' uptake of science courses at (higher) Leaving Certificate level?
- Yager and Weld (1999): How does the SS&C course affect students' achievement in concept, process, application, creativity, attitude and

worldview? How does the course affect this achievement for students of different gender and different ability?

- Zoller *et al.* (1990): What has been the impact of ST11 teaching on students' views of the relationship between Science, Technology and Society?

The following seven studies combined an evaluation of pupils' learning outcomes with a survey of their views of the nature of the context-based/STS approach:

- Ebenezer and Zoller (1993): What are the attitudes to science of students using the JSSP programme? What are students' perceptions of the JSSP programme they received and what are the aspects in learning and teaching affecting these perceptions?

- Huppert *et al.* (1992): What is the effect of the intervention [Human Health and Science] on pupils' understanding of science concept? What are pupils' attitudes towards the intervention?

- Nentwig *et al.* (2002): What is the effect of the intervention on students' understanding of chemical concepts, and on their ability to apply chemical knowledge to everyday contexts? How does the intervention affect learners' perceptions of the instructional quality?

- Ramsden (1997): What is the effect of context-based approach on pupils' understanding of key chemical ideas at 16+ and their career intentions? What are pupils' views on their chemistry lessons?

- Sutman and Bruce (1992): What is students' achievement in understanding chemical concepts at the end of the ChemCom intervention? What is the perceived effectiveness of ChemCom in achieving its learning objectives?

- Wierstra (1984): What is the effect of PLON on students' cognitive achievement, their affective and attitudinal outcomes, and what is the relationship between these effects? What is the effect on students' perceptions of their learning environment?

- Wierstra and Wubbels (1994): To what extent does the activity- and reality-centredness of the PLON curriculum affect affective and cognitive outcomes? To what extent is the activity- and reality-centredness of the PLON curriculum perceived as meaningful?

One study exclusively focused on pupils' views of their experiences of the context-based/STS approach differentiating for views of boys and girls:

- Ramsden (1992): What are pupils' reactions to the Salters course, and how do these reactions to the Salters course differ for boys and girls?

Most of the studies evaluate mature interventions as being four to six years old. However, the outcomes of three of the studies were intended to provide information for modification of the courses being evaluated. Huppert *et al.* (1992) and Nentwig *et al.* (2002) report on a pilot experience intended to refine course materials and teaching strategies. Greenlee and Lambert (1996) aimed at using the outcomes for the construction of a teachers' guide.

4.3.2 Methodological considerations

Table 4.3: Overview of studies' sampling methods and sample

Author(s), year	Features of study design	
	Sampling method	**Size and nature of actual sample**
Ebenezer and Zoller (1993)	All Grade 10 pupils in one district (convenient and representative); subset for interviews: 72 pupils in six schools stratified for positive, neutral and negative attitudes.	Questionnaires: 540 boys, 520 girls, 34 gender NI (69.9% of sample); interviews 13 boys, 19 girls. No details on characteristics of pupils.
Greenlee and Lambert (1996)	Selected from applicants to the programme. No sampling method.	24 girls. Volunteers. Average to above average ability girls. High interest in SEM from the start.
Huppert *et al.* (1992)	All pupils of one class piloting the intervention. No detail on selection method of school or class.	15 boys, 20 girls. Large proportion under achievers (15/35) in elective course.
Nentwig *et al.* (2002)	All pupils in eight classes of five trial schools taking ChiK. No detail how schools/classes were selected.	110 pupils. No details on gender breakdown. No details on characteristics of pupils.
Ramsden (1992)	All pupils in one class from six schools out of 100 trial schools, using eight or more Salters units.	65 boys, 59 girls. From mixed-ability, mixed sex teaching groups.
Ramsden (1997)	All pupils in year 11 in eight schools (four using Salters, four traditional approach). Predicted GCSE grades are used in matching pairs by ability.	168 pupils equally divided between experimental and control group. No details of gender distribution. Middle and higher ability pupils.
Reid and Skryabina (2002)	Cross-age analysis. Secondary data from five schools. No detail on how classes/schools were selected.	Experimental group (S3+S4 puplis): N=103+152 (69+87 boys, 43+65 girls); control group (S5/S6 pupils): N=96 (68 boys, 28 girls). No details on characteristics of pupils.
Smith and Bitner (1993)	Random selection of classes from five schools in Midwestern city (USA). No detail for selection method for schools.	Experimental group (ChemCom): N=60 (30 boys, 30 girls); control (GenChem): N=63 (32 boys, 31 girls). Urban pupils only. Experimental group non-science majors, control group science-majors.
Smith and Matthews (2000)	All Transition Year pupils of one school divided in experimental and control groups by researcher (staff member). Interviews with subset across ability range.	Experimental group: N=37 (18 boys, 19 girls); control group: N= 23 (12 boys, 11 girls). Interviews with 12 pupils (8 experimental, 4 control). Typically from deprived background with inclination towards early school leaving and low uptake of science at higher levels.
Sutman and Bruce (1992)	No details, but stratified for rural/urban schools.	3,700 pupils, no breakdown in ChemCom and control groups. 60% in pre-college, 40% non-college programmes. Half in urban, half in rural schools. Achievement below average.
Wierstra (1984)	Nine classes using PLON approach, six classes the traditional approach. No detail on method of selecting classes.	Experimental group: N=254 (189 boys, 65 girls); control group: N=144 (108 boys, 36 girls). Intermediate level pupils. No details on other pupil characteristics.
Wierstra and Wubbels (1994)	Eight classes using PLON approach, 16 classes the traditional approach. No detail on method of selecting classes.	Experimental group: N=209 pupils; control group N=355 pupils. No gender breakdown. Intermediate level pupils only. No details on other pupil characteristics.
Yager and Weld (1999)	All pupils from Grade 6–8 classes of teachers involved in SS&C intervention, each teaching at least one experimental class and one control class. Random allocation of classes to treatments by counsellors, representative of five districts involved.	5,270 (2673 boys, 2,557 girls) in experimental group, 1,320 (688 boys, 632 girls) in control group. Subset for career orientation: 364 (177 boys, 187 girls) in experimental group; 359 (179 boys, 180 girls) in control group. No details on characteristics of pupils.
Zoller *et al.* (1990)	Random selection from all 12th graders in six comprehensive schools in Vancouver	Experimental group: N=101 (41 boys, 60 girls) (37% of 473); control group N=276 (130 boys, 146

	area. No detail on selection method of schools.	girls) (19% of 1,496). five of six schools classify their pupils as 'non-academic'.

Sampling method

Table 4.3 provides an overview of the sampling methods, the actual sample size and the nature of the sample for the 14 studies. It shows that none of the studies used an explicit sampling frame, such as a roll of students in a school, the list of classes in a school, or the national or regional register of schools. Only one study (Yager and Weld, 1999) used all pupils in the specified sampling frame. Ebenezer and Zoller (1993) used all Grade 10 students using the JSSP curriculum in a district of British Columbia, selected for 'convenient access and representativeness of the schools in the district'. All other studies used a convenience sample for the identification of schools. They mostly used schools where access has been secured through their participation in the curriculum innovation being evaluated (for instance, Ramsden, 1992; Sutman and Bruce, 1992), or where a researcher has been on the staff as a teacher (for instance, Smith in Smith and Matthews, 2000). Such convenience sampling is probably realistic for research studies fitting in with practice.

Within schools, all but four studies used classes as the unit of sampling. The first exception was the study by Greenlee and Lambert (1996) whose sample consisted of volunteers. Secondly, Smith and Matthews (2000) were able to constitute their experimental and control class with a balance for gender. Within the identified classes, Ramsden (1997) carefully matched pairs of pupils across the experimental and control groups for predicted grades. Lastly, Zoller *et al.* (1990) made a random selection of pupils from identified classes. For the other studies, the allocation of classes to treatments was mostly unspecified, with some stratification mentioned for pupils from rural versus urban schools (Sutman and Bruce, 1992; Yager and Weld, 1999). All studies took the individual student as the unit of their evaluation. Thus they measured and reported the effect of interventions on the individual student's understanding, attitude or perceptions.

Sample size

As indicated by Table 4.3, three studies worked with samples of 20 to 60 pupils equivalent to one or two classes (Greenlee and Lambert, 1996; Huppert *et al.*, 1992; Smith and Matthews, 2000). Four studies used samples of between 100 and 200 pupils, generally comprising six to eight classes (Nentwig *et al.*, 2002; Ramsden, 1992; Ramsden, 1997; Smith and Bitner, 1993). A further four studies used samples of between 300 and 600 secondary school pupils (Reid and Skryabina, 2002; Wierstra, 1984; Wierstra and Wubbels, 1994; Zoller *et al.*, 1990). The remaining three studies used large samples of over 1,000 pupils (Ebenezer and Zoller, 1993; Sutman and Bruce, 1992; Yager and Weld, 1999), with the last study reporting the largest sample of 6590 pupils.

Characteristics of pupils in the samples

The majority of studies provide limited information about the characteristics of students in the sample. One could argue that such information is not necessary for very large samples as they may be assumed to be representative (for instance, Yager and Weld, 1999). However, some studies used clearly atypical small or smallish samples. As mentioned above, Greenlee and Lambert (1996) worked with volunteers. Other samples consisted of high-ability pupils (Greenlee and Lambert, 1996; Ramsden, 1997), low-ability pupils (Sutman and Bruce, 1992; Zoller *et al.*, 1990), poorly motivated pupils from lower socio-economic backgrounds (Smith and Matthews, 2000), or female pupils only (Greenlee and Lambert, 1996).

Table 4.4: Overview of studies' comparison groups and base-line data

Author(s), year	Features of study design	
	Comparison group	**Baseline data**
Ebenezer and Zoller (1993)	No comparison group, although data from earlier cohort are used as quasi control group. Comparison for gender.	No baseline data
Greenlee and Lambert (1996)	No comparison group for teaching approach	Yes, same pre- and post-test
Huppert *et al.* (1992)	No comparison group for teaching approach	No baseline data
Nentwig *et al.* (2002)	No comparison group for teaching approach	Yes, same test before and after each unit
Ramsden (1992)	No comparison group for teaching approach, but for gender	No baseline data
Ramsden (1997)	Experimental-control groups for teaching approach. Note control for ability.	No baseline data
Reid and Skryabina (2002)	Inferred experimental-control groups for teaching approach, comparing S4 (experimental) and S5/S6 (control) groups.	No baseline data
Smith and Bitner (1993)	Experimental-control groups for teaching approach. Discrepancy for ability between groups. Comparison for gender.	Yes. Pre- and post-test different but comparable
Smith and Matthews (2000)	Experimental-control groups for teaching approach	Yes, same pre- and post-tests
Sutman and Bruce (1992)	Experimental-control groups for teaching approach	No baseline data
Wierstra (1984)	Experimental-control groups for teaching approach. Comparison for gender.	No baseline data
Wierstra and Wubbels (1994)	Experimental-control groups for teaching approach. Comparison for gender.	Yes, same pre- and post-tests
Yager and Weld (1999)	Experimental-control groups for teaching approach. Comparison for gender and ability. Note control for teacher.	Yes, same pre- and post-tests
Zoller *et al.* (1990)	Experimental-control groups for teaching approach	None, but data from earlier cohort are used to argue that views of both groups are comparable.

Comparison/control of independent variable

Table 4.4 provides an overview of the controls and comparisons as part of the 14 studies. Seven studies used a quasi experimental design with an experimental group of pupils in classes using a context-based/STS approach and a control group in classes using a traditional teaching approach. A further study (Reid and Skryabina, 2002) used a cross-age analysis, but an experimental interpretation has been inferred by considering the S4 group as the experimental group (using an explicit context-based approach) and the older S5/S6 group as the experimental group (using a traditional approach). Ebenezer and Zoller (1993) reported only on a sample of pupils using an STS approach but used earlier reported data on a cohort using a non-STS approach as a quasi-control group. Four (generally small) studies used no control group for teaching approach (Greenlee and Lambert, 1996; Huppert *et al.*, 1992; Nentwig *et al.*, 2002; Ramsden, 1992).

Pre-post data-collection of dependent variable

Table 4.4. indicates that six studies used a prospective design and collected pre-intervention data as baseline information. Only Smith and Bitner (1993) used different sets of items from an established standardised item pool (GALT) for their pre- and post-tests. The other six studies used identical instruments as pre- and post-tests. Although Zoller *et al.* (1990) did not collect pre-intervention data from their sample, they argue that mid-intervention data from a different cohort may be used as baseline data. The remaining six studies did not collect data before the intervention started.

Table 4.5: Overview of studies' data-collection methods

Author(s), year	Data-collection	
	Instruments	**Reliability and validity**
Ebenezer and Zoller (1993)	Questionnaire (partly provided) with 29 Likert scale items based on BC Science Assessment Instrument. Semi-structured interview (schedule provided)	Hoyt reliability value 0.89 for questionnaire items Same instrument used in previous study Interview prompts are very general and positive.
Greenlee and Lambert (1996)	Standardised Tennessee Self-Concept Scale (TSCS), and Children's Academic Intrinsic Motivation Inventory (CAIMI) Self-developed questionnaires probing the same areas and perceptions on the intervention (provided)	No details of reliability Some standardised measures used
Huppert *et al.* (1992)	Self-constructed 10 Likert scale items attitude questionnaire Self-constructed achievement test with 15 MC and three open items (not provided)	Alpha Cronbach test 0.79 (five positive, five negative items) Both instruments peer-validated by five teachers
Nentwig *et al.* (2002)	54-item questionnaire about perceptions of ChiK Achievement test for chemical concepts and problem solving ability (not provided)	No details of reliability nor validity Use of trial schools could introduce bias
Ramsden (1992)	Self-developed questionnaire (provided): six item five-point Likert scale and one open question	No details of reliability Standard short descriptions of units provided as reference Use of trial schools could introduce bias
Ramsden (1997)	Eight choice and justification diagnostic items from previous study by Barker (example provided) and open questions for attitude (not provided)	No details of reliability Same instrument used in previous study
Reid and Skryabina (2002)	Osgood-type questionnaire (provided) statements with adjective antonyms for attitude to science lessons, self-perception and career intentions	No details of reliability Peer-validation Authors acknowledge potential bias due to self-selection in older samples

Author(s), year	Data-collection	
	Instruments	Reliability and validity
Smith and Bitner (1993)	12-item Group Assessment of Logical Thinking (GALT) test, with choice+justification responses	GALT test has well-established reliability and validity. GALT test items have 0.85 reliability with Piagentian interview tasks.
Smith and Matthews (2000)	25-item Likert-scale questionnaire on attitudes to science and career intentions 21-item similar questionnaire on actual and ideal learning environment (examples provided) Semi-structure interviews (schedule provided)	No detail of reliability Some items are modified from 'Students' Attitudes to Science' instrument.
Sutman and Bruce (1992)	11-item Likert-scale Opinionnaire Achievement test has (concept:application= 2:1) MC items only (not provided)	Reliability factor 0.80 (unclear how calculated) Achievement test: extensive peer-validation, piloted with large ability range Application responses correlated to Longeot test (corr. 0.4). Test published in teachers' guide may introduce bias
Wierstra (1984)	12-item Likert-scale attitude questionnaire used in earlier study 10-item Individualised Environment Questionnaire probing views of actual and preferred learning environment Achievement from PLON and standard exams	Actual learning items Alpha coeff. 0.72, preferred learning items Alpha coeff. 0.64 Inter-test correlation of 0.83 Same instrument used in previous study.
Wierstra and Wubbels (1994)	Achievement test with 19 multiple-choice items Attitude Likert-type questionnaires (examples provided)	Achievement test items Alpha 0.53 Attitude scales with alpha 0.77-0.85 No details of validity
Yager and Weld (1999)	Concept-specific tests for understanding of concept and application Generic Likert-type questionnaires for pupils' views on process, creativity, attitude and worldview from Iowa Assessment Package (IAP) Questionnaire on career intention with five Likert-scale items	Test-retest reliability of all Likert scale test of 0.88-0.92 Concept tests were topic tests provided by national panels, all peer-validated. Authors note that using IAP may introduce bias towards control group
Zoller *et al.* (1990)	Four-item questionnaire from VOSTS inventory form CDN with choice and justification responses (provided in full)	VOSTS inventory extensively tested for reliability and validity No justification of the low number of questionnaire items

Data-collection instruments

Table 4.5 summarises the nature of the data-collection instruments used in the 14 studies and the attention given to reliability and validity of data-collection in each case. The table indicates that all studies used written instruments to collect data from pupils, and two studies (Ebenezer and Zoller, 1993; Smith and Matthews, 2000) used semi-structured interviews with a small subset of pupils as an additional data source. It is noticeable that information on attitudes was collected in all but two studies through five-point, Likert-type items. The first exception was the use of Osgood-type questionnaire items with semantic adjective antonyms by Reid and Skryabina (2002). The second exception was Ramsden (1997), who used open-ended items to collect views on enjoyment of chemistry lessons and career intentions. The format of the items of the achievement tests used in seven studies was more varied. Two studies used multiple-choice questions (Sutman and Bruce, 1992; Wiersta and Wubbels, 1994), one study used multiple-choice questions and justification diagnostic items (Ramsden, 1997), and Huppert *et al.* (1992) used a combination of multiple-choice questions and open questions. The item format used in three studies was unclear (Nentwig *et al.*, 2002; Wierstra, 1984; Yager and Weld, 1999).

Five studies used standard tools, all widely tested for reliability and validity of their responses. Ebenezer and Zoller (1993) used the Science Assessment Instrument developed by Bateson; Greenlee and Lambert (1996) used the Tennessee Self-Concept Scales (TSCS) and the Children's Academic Intrinsic Motivation Inventory (CAIMI); Smith and Bitner (1993) used the GALT test (Group Assessment for Logical Thinking); Yager and Weld (1999) used the Iowa Assessment Package (IAP); and Zoller *et al.* (1990) selected items from VOSTS (Views on Science, Technology and Society). Another four studies used instruments they, or other researchers, had used in previous studies (Ramsden, 1997; Smith and Matthews, 2000; Wierstra, 1984; Wierstra and Wubbels, 1994). The remaining five studies used self-constructed questionnaires, although the item banks for testing the various aspects of attitude to science widely available in the research literature have often been used as starting points.

Reliability and validity of data-collection

Several studies provide Alpha Cronbach reliability coefficients for their research instruments (Huppert *et al.* (1992): 0.79; Sutman and Bruce (1992): 0.80; Wierstra (1984): 0.64-0.72 for various tests; Wierstra and Wubbels (1994): 0.77-0.85 for various tests). Ebenezer and Zoller (1993) report a Hoyt reliability value of 0.89. Four studies with some self-constructed instruments do not report any reliability measure and thus do not allow any judgements on their reliability (Nentwig *et al.*, 2002; Ramsden, 1992; Reid and Skryabina, 2002; Smith and Matthews, 2000).

More detail is provided on the validity of the studies. Self-constructed test items were peer-validated for content validity in three studies (Huppert *et al.*, 1992; Reid and Skryabina, 2002; Sutman and Bruce, 1992). No validation was reported for self-constructed test-items in three other studies (Nentwig *et al.*, 2002; Ramsden, 1992; Smith and Matthews, 2000).

The nature of evaluations of interventions, especially of evaluations of 'young' interventions using volunteer schools and teachers, may introduce some bias in favour of the intervention since participating teachers and schools are, by virtue of the fact that they have volunteered to participate, positive towards the intervention. This may affect the validity of the findings of studies by Nentwig *et al.* (2002) and Ramsden (1992).

Piloting data-collection instruments as a strategy for increasing validity, especially appropriate for self-constructed instruments, is under-used or under-reported. Only one study (Sutman and Bruce, 1992) made mention of this strategy. It should be noted that the use of instruments, developed and used previously for samples similar to those of the studies, could be considered as a pilot validation.

Table 4.6: Overview of studies' data analysis methods

Authors, year	Data analysis	
	Reliability	**Validity**
Ebenezer and Zoller (1993)	Scores aggregated in positive, neutral and negative categories, and frequencies compared between items Unreported triangulation of questionnaire and interview responses	Standard statistics used Not aggregating Likert responses Much of the gender breakdown is implausible
Greenlee and Lambert (1996)	Mean and SD calculated T-tests for significant differences between pre- and post-test No stats	Standard statistics used Aggregating Likert responses Multi-instrument data increases validity
Huppert *et al.* (1992)	Mean and SD calculated per Likert item Scores compared between items (no significance testing) Means for achievement scores compared with previous year marks No data	Standard statistics applied but lacks meaning due to lack of controls
Nentwig *et al.* (2002)	Factor analysis identifies five factors T-test for pre-post differences for whole sample and by gender and ability (based on previous year grade) No data	α-ratings of 0.74 to 0.86 for factors identified No detail on validity
Ramsden (1992)	Means (no SD) calculated per gender per item per unit Testing significance by χ^2 method and critical ratio analysis	Standard statistics used Not aggregating Likert responses
Ramsden (1997)	Means and SD calculated per school and whole samples χ^2 significance testing and critical ratio analysis No data for open responses	Elaboration of pre-tested coding scheme Standard statistics used No blind coding of open responses
Reid and Skryabina (2002)	Scores aggregated in positive, neutral and negative categories, and frequencies compared per item for age groups, χ^2 significance testing	Standard statistics used Not aggregating Osgood responses
Smith and Bitner (1993)	Scores aggregated in formal, transitional and concrete reasoning levels, and frequencies compared for groups and gender, χ^2significance testing Mean and SD calculated, groups' post-tests compared with pre-test as covariate	Standard statistics used No inter-scorer agreement reported The findings consistently contradict the conclusions
Smith and Matthews (2000)	Responses for each item classified in positive, neutral, negative and frequencies calculated for both groups T-test for significant differences for groups and genders No triangulation of questionnaire and interview responses No data	Standard statistics used Aggregating Likert responses
Sutman and Bruce (1992)	Mean discrepancy value calculated for Likert responses using Provus Discrepancy Model (values 0.38-1.08) Unclear significance test for achievement differences for groups No data	Aggregating Likert responses
Wierstra (1984)	Responses are analysed by correlations, T-tests ANOVA and MANOVA	Standard statistics used Solid exploration of correlations between tests Aggregating Likert responses
Wierstra and Wubbels (1994)	Responses are analysed by correlations, T-tests, ANOVA and MANOVA	Standard statistics used Solid exploration of correlations between tests Aggregating Likert responses

Authors, year	Data analysis	
	Reliability	Validity
Yager and Weld (1999)	'Improvement rates' constructed for concept and application achievement scores T-tests for testing significant differences between groups, gender and ability ANOVA for testing significant differences of aggregated post-test scores for questionnaires compared for groups, gender and ability with pre-test as covariate	Standard statistics used Aggregating Likert responses Use of improvement rates, and comparison of pupils from same teacher increases validity
Zoller *et al.* (1990)	Response profiles developed by researchers from data Significance testing of differences between groups and genders by χ^2 method	Standard statistics used No independent co-construction of responses profiles

Data-analysis methods

Table 4.6 provides an overview of the analysis methods adopted by the 14 studies, and the ways of securing reliability and validity in data analysis. All studies report the use of some form of statistical analysis which, if done appropriately, provides a measure of reliability for the analysis.

Six studies (Ebenezer and Zoller, 1993; Ramsden, 1992; Ramsden, 1997; Reid and Skryabina, 2002; Smith and Bitner, 1993; Zoller *et al.*, 1990) report grouping responses to individual Likert-scale items and use χ^2 methods for the frequencies of each group to identify significance levels of differences between pupil groups for a particular item, or for the whole sample between items. For instance, Ebenezer and Zoller (1993) group pupil attitude responses to Likert-type items as positive, negative or neutral, and use frequencies to establish significant differences in attitude to the issue raised in the item.

Five studies (Ebenezer and Zoller, 1993**;** Greenlee and Lambert, 1996; Nentwig *et al.*, 2002; Ramsden, 1997; Smith and Mathews, 2000) used t-tests for identifying the significance of differences between treatment and control groups or between genders. All these studies, apart from Ramsden (1997), compare accumulated attitude scores of Likert-type items (with problematic validity). Ramsden (1997) in contrast used accumulated scores (validly) for comparing achievement scores.

Four studies (Smith and Bitner, 1993; Wierstra, 1984; Wierstra and Wubbels, 1994; Yager and Weld, 1999) used analysis of variance (ANOVA) methods for identifying the significance of differences in the performance of various groups (for instance, experimental versus control groups) after an intervention. The last three of these studies use accumulated attitude scores for Likert-type responses as the basis for comparison. Usually, significance levels for differences in post-intervention scores were compared with pre-intervention scares as a covariate. The reports of these studies provided details of group sizes, mean scores and standard deviations, F-values and p-values.

Two studies (Wierstra, 1984; Wierstra and Wubbels, 1994), used MANOVA methods. For instance, Wierstra (1984) analyses aggregated responses for boys and girls within the treatment group to Likert-scale type items on attitude towards the intervention by MANOVA to show significant differences between genders with regard to cognitive and affective outcomes. P-values were provided. Additional univariate tests showed higher scores for boys.

Reliability of data analysis

Since all the studies used some form of statistics to some extent, reliability issues have been taken care of. However, some studies had opportunities for additional basic reliability checks that were not pursued. For instance, Zoller *et al.* (1990) mentioned that response profiles were constructed by the researchers, but did not mention independent construction of these profiles, nor inter-researcher agreement. Equally, Ramsden (1997) did not report independent classification of responses to open-ended questions in her questionnaire by more than one researcher. Greenlee and Lambert (1996) collected data on pupils' self concept and motivation by two different written instruments, but no information was provided for individual pupils about the reliability of data collected through each instrument.

Validity of data analysis

Eight of the studies (Ebenezer and Zoller, 1993; Huppert *et al.*, 1992; Nentwig *et al.*, 2002; Smith and Matthew, 2000; Sutman and Bruce, 1992; Wierstra, 1984; Wierstra and Wubbels, 1994; Yager and Weld, 1999) dealt with Likert-type responses as aggregates, assuming that the five or six options are equidistant, and that similar (dis)agreements for different Likert statements carry the same weight. This is not necessarily the case. Weighing responses to Likert-type items (transforming them from discrete to continuous data) and aggregating these weighted responses for different Likert-scale items may affect the validity negatively.

Multi-source triangulation was a method used for data-collection in two studies but its usefulness for validity is rarely highlighted. For instance, Ebenezer and Zoller (1993), and Smith and Matthews (2000) collected interview data from a subset of the pupils in their sample. Although the selection of the interview sample has been based on the nature of the questionnaire responses, there is no report of the extent to which oral data coincide with written data of the same individuals, and of the extent to which interpretations of the written data are born out by the interview responses.

It should be noted that the data presented by Smith and Bitner (1993) consistently contradict their conclusions, so, for the purposes of this review, we have used their data and not their conclusions. Equally, several of the gender breakdowns (only a minor portion in the actual study) in the data by Ebenezer and Zoller (1993) are seen as implausible, and thus the synthesis for this review will disregard any conclusions based on these data.

Weighted evidence

Taking account of the different methodological aspects above, the quality of the 14 studies can be summarised as in Table 4.7. These quality weightings have been made against the declared aims, hypotheses and research questions of the respective studies. The weight of evidence A is that concluded in answer to question M.11 at the end of the data-extraction exercise, *'Taking account of all quality assessment issues, can the study findings be trusted in answering the study question(s)?'*

Table 4.7: Quality of the studies (weight of evidence A)

Study	Quality of the study (WoE A)
Ebenezer and Zoller (1993)	Medium
Greenlee and Lambert (1996)	Medium low
Huppert *et al.* (1992)	Low
Nentwig *et al.* (2002)	Low
Ramsden (1992)	Medium
Ramsden (1997)	High
Reid and Skryabina (2002)	Medium high
Smith and Bitner (1993)	Medium low
Smith and Matthews (2000)	Medium low
Sutman and Bruce (1992)	Medium low
Wierstra (1984)	Medium
Wierstra and Wubbels (1994)	High
Yager and Weld (1999)	High
Zoller *et al.* (1990)	Medium low

4.3.3 Appropriateness of the studies' research design for this review (category B)

This section of the report provides an overview of the appropriateness of the research design for the in-depth review question for the 14 studies in the review. This will provide the weight of evidence category B (WoE B).

The in-depth review questions are as follows:

What is the evidence from evaluative studies of the effect of context-based or STS courses on the attitude to science and/or the understanding of science ideas of boys and girls in the 11 to 16 age range?

What is the evidence from evaluative studies of the effect of context-based or STS courses on the attitude to science and/or on the understanding of science ideas of lower-ability pupils in the 11 to 16 age range?

The weights of evidence will therefore be presented under two different sub-headings: firstly, 12 studies will be analysed for appropriateness of the study design for measuring effects on gender (for the first in-depth review question); and, separately, five studies will be analysed for appropriateness of the study design for measuring effects on lower-ability pupils (for the second in-depth review question). Further details of the categories used to determine the review-specific WoE B are given in appendices 2.5.1, 2.5.2, 2.5.3.

Gender: appropriateness of research design

Research designs are weighted according to one precondition (the evaluative component of the study needs to apply to the effect of context-based/STS approach on boys and girls). In addition, five design aspects are graded: the appropriateness of the sampling method and sample size, the appropriateness of the comparison/control for the independent variable (gender), the prospectiveness of the dependent variable (attitude to science or understanding of science ideas), the appropriateness of the data-collection, and the appropriateness of the analysis methods. Table 4.8 provides a summary of these aspects for the 12 studies involved in the review on effects on gender. For each study, the second column indicates if effects on *attitude* to science (GA) or on *understanding* of science ideas (GU) are reported.

All studies were evaluations, but not all took gender as a major evaluative variable. Four studies (Ebenezer and Zoller, 1993; Nentwig *et al.*, 2002; Ramsden, 1997; Wierstra and Wubbels, 1994) evaluated the effect of a context-based/STS approach more generally, but did not take gender as a determining variable for the study's design. They reported on gender only to confirm that effects were gender independent. Consequently, these studies were given a low weight of evidence in terms of appropriateness of research design and are not discussed further in this section (4.3.3) because gender was not the focus of the design of the study. However, relevant findings will be discussed in the synthesis section (4.4).

Since the in-depth review intends to establish broadly generalisable evidence for the effects of context-based/STS approaches, a sampling method aimed at representativeness strengthens the weight of evidence for the findings of a study. All studies lack detail on claims of generalisibility (or lack of it) of their findings.

For the purposes of this review, a sample size of over 100 pupils, or four to five classes in treatment and control groups, is considered reasonable and a sample size of over 300 pupils will provide good confidence for generalising findings and conclusions. The three large studies used more than ten classes and are thus most appropriate for this in-depth review.

Table 4.8: Overview of appropriateness of research design for the in-depth review (GENDER)

Author(s), year	Review focus	Sample size/method for gender	Comparison for gender	Benchmark data for gender	Trustworthiness of data-collection for gender	Trustworthiness of data analysis for gender
Ebenezer and Zoller (1993)	GA	No evaluation of gender effect	-	-	-	-
Greenlee and Lambert (1996)	GA	Very small sample, with no sampling method	No control group (female sample only)	Pre- and post-intervention data for some measures	Some published tests, not all. Triangulation.	Standard statistics, but small sample and lack of control
Nentwig *et al.* (2002)	GA	No evaluation of gender effect	-	-	-	-
Ramsden (1992)	GA	Reasonable sample size, with good sampling method	Gender comparison in the design	No benchmark data for attitudes by gender	No reliability checks, summary of units improves validity	Standard statistics, no validity check for open responses
Ramsden (1997)	GA	No evaluation of gender effect	-	-	-	-
Reid and Skryabina (2002)	GA	Large sample, with no sampling method	Gender comparison in design	No benchmark data for attitudes by gender	No reliability checks. Peer-validation of instrument.	Standard statistics
Smith and Bitner (1993)	GU	Reasonable sample size, with clear sampling method	Gender comparison in design	Good pre-post collection of data for understanding	Some checks on validity/reliability, no control of ability	Data contradict the conclusions
Smith and Matthews (2000)	GA	Small sample Prospective allocation	Good gender comparison in design	Good pre-post collection of data for attitude	Adapted published tests, omitting half the variables	Standard statistics, but no values provided
Wierstra (1984)	GA	Large sample, with clear sampling method	Gender comparison in the findings, not the design	No pre-intervention data	Previously used attitude tests, with high α value	Standard statistics Correlations but no data provided
Wierstra (1984)	GU	Large sample, with clear sampling method	Gender comparison in the findings, not the design	No pre-intervention data	Standard PLON and traditional tests used	Standard statistics Significance tests but no data provided
Wierstra and Wubbels (1994)	GA	No evaluation of gender effect	-	-	-	-
Wierstra and Wubbels (1994)	GU	No evaluation of gender effect	-	-	-	-
Yager and Weld (1999)	GA	Very large sample, with careful sampling method	Good comparison for gender in design	Pre- and post-intervention data for attitude included	Solid checks on reliability/validity (published test, test-retest)	Solid checks on reliability/validiity (ANOVA) for attitude
Yager and Weld (1999)	GU	Very large sample, with careful sampling method	Good comparison for gender in design	Pre- and post-intervention data for understanding	Solid checks on reliability/validity (published test, test-retest)	Solid checks on reliability/validity (T-tests on improvement rates)
Zoller *et al.* (1990)	GA	Large sample No detail on sampling method	Good comparison for gender in design	No benchmark data for gender	Published test, but very few items	Reasonable reliability/validity checks, but no inter-coder rates

Although hardly any of the studies give detail of socio-economic, cultural and ability characteristics, the larger studies may be considered to include a cross-section of pupils. On the other hand, several smaller studies used atypical samples. The findings of these studies with unrepresentative samples would have limited generalisibility.

Although all study designs were included in this review, higher weight of evidence was accorded to those studies in the in-depth review with a control group as comparison. All but one of the remaining eight studies used groups of boys and girls as a basis for comparison in their design, the exception being Greenlee and Lambert (1996) with an all-female sample. Several of these studies carefully matched the experimental and control groups for teacher effect (Yager and Weld, 1999), and for prior knowledge levels (Smith and Bitner, 1993; Yager and Weld, 1999). In contrast, Smith and Bitner (1993) built in some unchecked covariance as their treatment group consists entirely of non-science majors, whereas the control group is made up of pupils intent on pursuing further science studies.

Studies with a prospective design measuring boys' and girls' attitude to science and/or their understanding of science ideas before and after an intervention are most appropriate for this in-depth review. Four studies (Greenlee and Lambert, 1996; Smith and Bitner, 1993; Smith and Matthews, 2000; Yager and Weld, 1999) fell into this category. The other four studies only collected post-intervention data.

Several studies used established test instruments developed for samples with similar characteristics, thus increasing the reliability and validity of the responses. This was particularly the case for studies collecting data on pupils' understanding of science concepts and application; all three studies (Smith and Bitner, 1993; Wierstra, 1984; Yager and Weld, 1999) used national examination items. Several attitude measures were drawn from established tests. For instance, Smith and Bitner (1993) used the GALT test for scientific reasoning. However, some studies (Ramsden, 1992; Reid and Skryabina, 2002; Smith and Matthews, 2000) construct their own tests for pupil attitudes, usually with few reported checks on their validity or reliability.

Most studies used standard statistical methods for establishing significance levels of differences between attitudes or achievement of boys and girls. In most of these studies, Likert-scale responses were scored and aggregated across items, thus introducing some doubt on the validity of the findings. The exceptions were the studies by Ramsden (1992) and Zoller *et al.* (1990) where within-item comparisons were conducted. Verification of the findings by Smith and Matthews (2000) and Wiersta (1884) was impossible as no data (only correlation coefficients or p-values) were provided. The conclusions by Smith and Bitner (1993) consistently seem to contradict the data presented.

Lower ability: appropriateness of research design

Research designs are weighted according to one precondition: the evaluative component of the study needs to apply to the effect of context-based/STS approach on lower-ability pupils. As explained above for gender, five additional design aspects are graded: the appropriateness of the sampling, the comparison for the independent variable (lower ability), the prospectiveness of the dependent variable (attitude to science or understanding of science ideas), the appropriateness of the data-collection, and the appropriateness of the analysis methods. Table 4.9 provides a summary of these aspects for the five studies involved in the review on effects on lower-ability pupils. For each study, the second column indicates if effects on *attitude* to science (LA) or on *understanding* of science ideas (LU) are reported.

Table 4.9: Overview of appropriateness of research design for the in-depth review (LOWER ABILITY)

Author(s), year	Review focus	Sample size/method for lower ability	Comparison/control group for lower ability	Benchmark data for lower ability	Trustworthiness of data-collection for lower ability	Trustworthiness of data analysis for lower ability
Huppert *et al.* (1992)	LA	Very small sample, with no sampling method	No control group (many underachievers)	No data on pre-intervention attitude	Peer-validation and reliability of 0.79	Standard statistics, but aggregates Likert scores
Huppert *et al.* (1992)	LU	Very small sample, with no sampling method	No control group (many underachievers)	Last year's biology grade taken as benchmark data	Peer-validation, no check on reliability	Standard statistics, but small sample and lack of control
Nentwig *et al.* (2002)	LA	No evaluation of effect on lower-ability pupils	-	-	-	-
Smith and Matthews (2000)	LA	Small sample size Prospective allocation	No control group for ability (many low ability)	Good pre-post collection of data for attitude	Adapted published tests, omitting half the variables	Standard statistics, but no values provided
Sutman and Bruce (1992)	LU	Very large sample, with no sampling method	No control group for ability (many low ability)	No benchmark data on understanding	Solid checks on instrument validity	No information on analysis method
Yager and Weld (1999)	LA	Large sample for low ability (bottom third)	Good comparison for ability only in findings	Pre- and post-intervention data for attitude	Solid checks on reliability/validty (published test, test-retest)	Solid checks on reliability/validity (ANOVA) for attitude
Yager and Weld (1999)	LU	Large sample for low ability (bottom third)	Good comparison for ability only in findings	Pre- and post-intervention data for understanding	Solid checks on reliability/validity (published test, test-retest)	Solid checks on reliability/validity (T-tests on improvement rates)

The evaluation by Nentwig *et al.* (2002) reports on the effects of teaching approaches on lower-ability pupils, but it did not take lower ability as a determining variable for the study's design. Consequently, no further details of the design of this study have been presented as it is inappropriate for this review.

Two of the remaining four studies used very small samples (Huppert *et al.*, 1992; Smith and Matthews, 2000), whereas the other two used samples of over 1,000 pupils. In addition, Yager and Weld (1999) report a careful sample selection method, thus providing good confidence for generalising findings and conclusions.

Only one study (Yager and Weld, 1999) explicitly compared effects for lower- and higher-ability pupils. They constructed the lower- and higher-ability groups within their sample as the one-third bottom achievers and the one-third top achievers respectively. The tremendous size of the sample suggests that potential validity problems with this internal stratification sampling strategy may have been avoided. The other three studies were included in this review as they deal with pre-selected and coherent lower-ability groups.

Studies with a prospective design measuring attitude to science and/or understanding of science ideas of low-ability pupils before and after an intervention are most appropriate for this in-depth review. Two studies collected such benchmark data (Smith and Mathews, 2000; Yager and Weld, 1999). Huppert *et al.* (1992) used previous year's biology grades as proxy for benchmark data for understanding of science ideas. The other study (Sutman and Bruce, 1992) only collected post-intervention data.

Several studies (Huppert *et al.*, 1992; Sutman and Bruce, 1992) used extensive peer-validation of self-constructed test instruments, thus increasing the validity of the responses. Yager and Weld (1999) used established tests for data-collection: national examination items for testing understanding, and the Iowa Assessment Package for attitude.

Most studies used standard statistical methods for establishing significance levels for changes in attitudes or achievement for lower-ability pupils. Likert-scale responses were scored and aggregated across items in two studies (Huppert *et al.*, 1992; Yager and Weld, 1999), thus introducing some doubt on the validity of the findings. Verification of the findings by Smith and Matthews (2000) and Sutman and Bruce (1992) was impossible as no data (only p-values) are provided.

Weighted evidence (category B)

Taking account of the different methodological aspects above, the quality of the 14 studies can be summarised as in Table 4.10. These quality weightings have been made against the appropriateness of the study design for the in-depth review question.

Table 4.10: Appropriateness of the study design (weight of evidence B)

Author(s), year		Evaluation aspect			
		Gender		Lower ability	
		Attitude (GA)	Understanding (GU)	Attitude (LA)	Understanding (LU)
Ebenezer and Zoller (1993)		L			
Greenlee and Lambert (1996)		ML			
Huppert *et al.* (1992)				L	L
Nentwig *et al.* (2002)		L		L	
Ramsden (1992)		M			
Ramsden (1997)		L			
Reid and Skryabina (2002)		M			
Smith and Bitner (1993)			MH		
Smith and Matthews (2000)		M		ML	
Sutman and Bruce (1992)					ML
Wierstra (1984)		M	M		
Wierstra and Wubbels (1994)		L	L		
Yager and Weld (1999)		H	H	H	H
Zoller *et al.* (1990)		M			
Total frequencies	*H*	*1*	*1*	*1*	*1*
	MH	*0*	*1*	*0*	*0*
	M	*5*	*1*	*0*	*0*
	ML	*1*	*0*	*1*	*1*
	L	*4*	*1*	*2*	*1*
	Total	***11***	***4***	***4***	***3***

L = Low, ML = Medium low, M = Medium, MH = Medium high, H = High

4.3.4 Relevance of the study focus for the in-depth review (category C)

This section of the report provides an overview of the appropriateness of the research focus for the in-depth review questions for each of the 14 studies. This will provide the weight of evidence category C (WoE C).

Again, the weight of evidence will be presented under two different subheadings: firstly, 12 studies will be analysed for appropriateness of the focus of the studies for measuring effects on gender (for the first in-depth review question); and separately, five studies will be analysed for appropriateness for measuring effects on lower-ability pupils (for the second in-depth review question). Further details of the categories used to determine the review-specific weight of evidence C are given in appendices 2.5.1, 2.5.2 and 2.5.3.

Gender: appropriateness of research focus

The relevance of the focus of 12 studies will be weighted according to five aspects: the nature and specificity of the independent variable (gender), the nature and breadth of the measures of the dependent variable (attitude to science or understanding of science ideas), and the representativeness of the research context. Table 4.11 provides a summary of these aspects for the 12 studies involved in the review on effects on gender. For each study, the second column indicates if effects on *attitude* to science (GA) or on *understanding* of science ideas (GU) are reported.

Half the studies had treatment and/or control groups well balanced for gender. Two studies had considerably more boys than girls in their treatment groups (Wierstra, 1984; Zoller *et al.*, 1990), and three studies did not provide a breakdown for gender in their samples (Nentwig *et al.*, 2002; Ramsden, 1997; Wierstra and Wubbels, 1994). Greenlee and Lambert (1996) worked with an all-female sample.

In the Ramsden (1992) study, gender was the sole independent variable, whereas gender was one of the variables in studies by Reid and Skryabina (2002), Smith and Bitner (1993), Smith and Matthews (2000), Yager and Weld (1999), and Zoller *et al.* (1990). Mostly, the other independent variables included the nature of the teaching approach and/or pupil ability. In the other studies, gender was not treated as an explicit variable.

Table 4.11: Overview of the relevance of the study focus for in-depth review (Gender)

Authors, year	Review focus	Representativeness of sample for gender	Nature of gender as variable	Appropriateness of measures for A/U	Breadth of measures for A/U	Representativeness of classroom situation
Ebenezer and Zoller (1993)	GA	Balanced gender ratio for written and oral data	Gender is not an explicit focus of evaluation	Some items test attitude, some report activities	Limited test (four issues) for attitude to school science	Representative for learners in schools
Greenlee and Lambert (1996)	GA	Volunteer, able girls with science interest	Gender is a major focus of evaluation/intervention	Appropriate measures for attitude	Broad range of attitudes explored	Three-week residential course not representative of class
Nentwig *et al.* (2002)	GA	No data on gender breakdown	Gender is not an explicit focus of the evaluation	Seemingly good attitude measures (little detail)	Good breadth for views on instruction only	Representative of pilot classrooms
Ramsden (1992)	GA	Balanced gender ratio in intact classes	Gender is the explicit variable	Highly appropriate (Likert-scale and elaboration)	Broad range of attitudes explored	Highly representative (mixed ability, mixed sex)
Ramsden (1997)	GA	No data on gender breakdown	Gender is not an explicit focus of the evaluation	Open items on enjoyment good (incomplete report)	Narrow range of attitudes (enjoyment, careers)	Representative of high ability classrooms
Reid and Skryabina (2002)	GA	Reasonable gender ratios in intact classes	Gender is one of the explicit variables	Highly appropriate for measuring attitudes	Broad range of attitudes explored	Standard curriculum representative of classes
Smith and Bitner (1993)	GU	Balanced gender ratio in intact classes	Gender is one of the explicit variables	Measures good in testing science understanding	Breadth of understanding is limited (reasoning only)	Representative of classroom situation
Smith and Matthews (2000)	GA	Balanced gender ratio in intact classes	Gender is one of the explicit variables	No measure of views of science and environment	Reasonable range of attitudes measured	Prevalence of low achieving pupils
Wierstra (1984)	GA	Gender balance not ideal (boys: girls = 3:1)	Gender was one variable (also intervention, teacher)	Used Likert-scale measure for attitude (no detail)	Good breadth of attitudes	Highly representative of intermediate level classes
Wierstra (1984)	GU	Gender balance not ideal (boys: girls = 3:1)	Gender was one variable (also intervention, teacher)	Test from each curriculum for understanding	Good breadth of understanding	Highly representative of intermediate level classes
Wierstra and Wubbels (1994)	GA	No data on gender breakdown	Gender is seen as co-variable to be controlled	Little detail given on attitude measure	Little detail for breadth of attitude measure	Unclear
Wierstra and Wubbels (1994)	GU	No data on gender breakdown	Gender is seen as co-variable to be controlled	Little detail for measure of understanding	Reasonable breadth of measure of understanding	Unclear
Yager and Weld (1999)	GA	Balanced gender ratio in intact classes	Gender is one of the explicit variables	Attitude measures from Iowa Assessment Package (IAP)	Broad attitude measures (covers two of six domains)	Highly representative of classrooms
Yager and Weld (1999)	GU	Balanced gender ratio in intact classes	Gender is one of the explicit variables	Understanding measures from national test bank	Tests of understanding cover concept and application	Highly representative of classrooms
Zoller *et al.* (1990)	GA	Gender ratio in exp group a little unbalanced	Gender is one of the explicit variables	Constructed profiles good, but based on few items	Attitudes seriously limited to STS relationships	'Non-academic' pupils taught by non-specialists

Understanding of science ideas was measured by validated multiple-choice test items. For measuring attitude, Likert-type items were used by most as is commonly accepted as appropriate. Ramsden (1992) supplemented these with an open question for elaboration, thus providing an additional opportunity for validation of the responses. Whereas attitude scales should invite views and opinions, Ebenezer and Zoller (1993) and Smith and Matthews (2000) included a set of items soliciting reports of (perceived) actual learning environment, without asking for views of the preferred environment. Ramsden (1997) used open questions for measuring enjoyment but her report is incomplete.

The range of understanding being measured depended on the concepts that had been taught during the intervention. Yager and Weld (1999) specifically tested understanding of science concepts *and* the ability to apply these concepts. The range of attitudes being measured varied from attitudes to school science teaching (Ebenezer and Zoller, 1993; Nentwig *et al.*, 2002; Ramsden, 1992; Smith and Matthews, 2000), attitudes to the learning environment (Ebenezer and Zoller, 1993; Reid and Skryabina, 2002; Smith and Matthews, 2000; Wierstra and Wubbels, 1994; Yager and Weld, 1999), attitudes to science as a subject (Wierstra and Wubbels, 1994; Zoller *et al.*, 1990), attitudes to further studying science (Greenlee and Lambert, 1996; Ramsden, 1997; Reid and Skryabina, 2002; Smith and Matthews, 2000; Yager and Weld; 1999), enjoyment of the intervention (Greenlee and Lambert, 1996; Ramsden, 1992; Ramsden, 1997; Yager and Weld, 1999), usefulness of the intervention (Nentwig *et al.*, 2002; Ramsden, 1992; Reid and Skryabina, 2002; Wierstra and Wubbels, 1994), and self-concept as a science learner (Greenlee and Lambert, 1996; Nentwig *et al.*, 2002; Reid and Skryabina, 2002).

Half the studies worked with mixed-ability and mixed-gender intact classrooms (Ramsden (1992) used this as a sampling condition), and thus provide confidence for generalisation of their findings. Some studies used samples with a limited ability range. Greenlee and Lambert (1996) used a three-week residential course as their intervention, rather than a standard classroom curriculum implementation. Wierstra (1984) reports little about the setting of his study.

Lower ability: appropriateness of research focus

The relevance of the focus of five studies will be weighted according to five aspects: the nature and specificity of the independent variable (lower ability), the nature and breadth of the measures of the dependent variable (attitude to science or understanding of science ideas), and the representativeness of the research context. Table 4.12 provides a summary of these aspects for the five studies involved in the review on effects on lower-ability pupils. For each study, the second column indicates if effects on *attitude* to science (LA) or on *understanding* of science ideas (LU) are reported.

It is difficult to be sure about the nature of 'lower ability' in the five studies. Only Yager and Weld (1999) had sizeable and balanced groups of lower-ability (and higher-ability) pupils as their ability samples. They selected these groups on the basis of the distribution of achievement scores within the total sample. Huppert *et al.* (1992) report that just over 40% of their sample were lower-ability pupils. Sutman and Bruce (1992) make a general statement that '40% are non-science majors' (which may or may not be identical to lower ability), whereas Smith and Matthews (2000) describe their sample as a mixture of disaffected and underachieving. Nentwig *et al.* (2002) do not provide a breakdown for what they report as lower-ability pupils within their sample.

In four of the five studies, lower ability (or ability) was one of the various independent variable considered. Only Nentwig *et al.* (2002) did not set out to look at the effects of low ability.

Table 4.12: Overview of the relevance of the study focus for the in-depth review (Lower ability)

Author(s), year	Review focus	Representativeness of sample for lower ability	Nature of lower ability as variable	Appropriateness of measures for A/U	Breadth of measures for A/U	Representativeness of classroom situation
Huppert *et al.* (1992)	LA	No detail of type of low ability, only percentage	Lower ability is one of the variables.	Reasonable measure of attitude (no examples)	Attitude measures limited to course objectives	Representative of classroom in elective course only
Huppert *et al.* (1992)	LU	No detail of type of low ability, only percentage	Lower ability is one of the variables.	Measure of understanding unclear (no examples)	Reasonable measures of understanding	Representative of classroom in elective course only
Nentwig *et al.* (2002)	LA	No data on ability breakdown	Low ability is not one of the explicit variables.	Seemingly good attitude measure (little detail)	Good breadth for views on instruction only	Representative of pilot classrooms
Smith and Matthews (2000)	LA	Detailed information on low ability/disaffection	Lower ability is one of the variables.	No measure of views of science and environment	Reasonable range of attitudes measured	Prevalence of low-achieving pupils
Sutman and Bruce (1992)	LU	Majority non-science majors, not lower ability	Low ability/non-science major is discrete element.	Highly scrutinised test for achievement	Understanding of concept and application (lacks detail)	Highly representative for classrooms
Yager and Weld (1999)	LA	Equal proportion high and low ability (1/3)	Ability is one of the explicit variables.	Attitude measures from IAP	Broad attitude measures (covers two of six domains)	Highly representative of classrooms
Yager and Weld (1999)	LU	Equal proportion high and low ability (1/3)	Ability is one of the explicit variables.	Understanding measures from national test bank	Tests of understanding cover concept and application	Highly representative of classrooms

Understanding of science ideas was measured by validated multiple-choice test items (Sutman and Bruce, 1992) or by the use of standardised tests (Yager and Weld, 1999). For measuring attitude, Likert-scale items were used by most, as is commonly accepted as appropriate. The validity and reliability of data-collection methods used by Huppert *et al.* (1992), Nentwig *et al.* (2002), and Smith and Matthews (2000) are difficult to judge due to lack of details about their instrument.

The range of understanding and attitude being measured is the same as for gender effects reported above.

The setting of the study by Yager and Weld (1999) is highly representative. The setting of the study by Nentwig *et al.* (2002) is only marginally less so since it involves intact classes selected for piloting an innovation. However, the settings of the other studies are (unsurprisingly) unrepresentative as they draw on samples of disaffected pupils, or evaluate elective classes or courses for non-science majors.

Weighted evidence (Category C)

Taking account of the different aspects above, the quality of the 14 studies can be summarised as in Table 4.13. These quality weightings have been made against the relevance of the focus of the study for the in-depth review question.

Table 4.13: Relevance of study's focus (weight of evidence C)

		Evaluation aspect			
		Gender		Lower ability	
		Attitude (GA)	Understanding (GU)	Attitude (LA)	Understanding (LU)
Ebenezer and Zoller (1993)		M			
Greenlee and Lambert (1996)		M			
Huppert *et al.* (1992)				M	M
Nentwig *et al.* (2002)		M		M	
Ramsden (1992)		H			
Ramsden (1997)		L			
Reid and Skryabina (2002)		H			
Smith and Bitner (1993)			MH		
Smith and Matthews (2000)		M		M	
Sutman and Bruce (1992)					MH
Wierstra (1984)		MH	MH		
Wierstra and Wubbels (1994)		ML	ML		
Yager and Weld (1999)		H	H	M	MH
Zoller *et al.* (1990)		M			
Total frequencies	*H*	3	1	0	0
	MH	1	2	0	2
	M	5	0	4	1
	ML	1	1	0	0
	L	1	0	0	0
	Total	***11***	***4***	***4***	***3***

L = Low, ML = Medium low, M = Medium, MH = Medium high, H = High

4.3.5 Overall weighting

Studies were given a rating on a five-point scale in each of the categories of weight of evidence (WoE): that is, the quality of the study in its own right (WoE A), the appropriateness of the study's design for this specific in-depth review question (WoE B), and the relevance of the focus of the study for this in-depth review question (WoE C). These weights of evidence, together with the overall weight for each study (WoE D), are summarised in Table 4.14.

Table 4.14: Overall weight of evidence (WoE)

		Evaluation aspect			
		Gender		Lower ability	
		Attitude (GA)	Understanding (GU)	Attitude (LA)	Understanding (LU)
Ebenezer and Zoller (1993)		ML*			
Greenlee and Lambert (1996)		ML			
Huppert *et al.* (1992)				ML	ML
Nentwig *et al.* (2002)		ML*		ML*	
Ramsden (1992)		MH			
Ramsden (1997)		ML*			
Reid and Skryabina (2002)		MH			
Smith and Bitner (1993)			M		
Smith and Matthews (2000)		M		ML	
Sutman and Bruce (1992)					M
Wierstra (1984)		M	M		
Wierstra and Wubbels (1994)		ML*	ML*		
Yager and Weld (1999)		H	H	MH	H
Zoller *et al.* (1990)		M			
Total frequencies	*H*	1	1	0	1
	MH	2	0	1	0
	M	3	2	0	1
	ML	5	1	3	1
	L	0	0	0	0
	Total	***11***	***4***	***4***	***3***

L = Low, ML = Medium low, M = Medium, MH = Medium high, H = High
* Study not designed to evaluate effects of gender or lower ability

Thus just over half the studies were deemed to have an overall weight of evidence of medium or better, with the remainder having lower overall weights of evidence.

Appendix 4.1 provides summary tables of the 14 studies included in the in-depth review. These tables are based on the information gathered and judgements reached in the data-extraction of the studies. Where a concise summary was included in the studies, the key conclusions in relation to understanding and attitude have been presented in the author's own words.

4.3.6 Summary of the evidence

Fourteen studies met the inclusion criteria for the in-depth review questions. Twelve studies have been included in the synthesis of evidence on gender, 11 for gender and attitude (GA), and four for gender and understanding (GU). Five studies have been included in the synthesis of evidence on lower ability, four on lower ability and attitude (LA), and three on lower ability and understanding (LU). Tables 4.15 to 4.18 summarise the weights of evidence assigned to the studies for each of these facets of the review.

Table 4.15: Weights of evidence for gender and attitude (GA)

Study	Weight of evidence A	Weight of evidence B	Weight of evidence C	Weight of evidence D
Ebenezer and Zoller (1993)	Medium	Low	Medium	Medium low*
Greenlee and Lambert (1996)	Medium low	Medium low	Medium	Medium low
Nentwig *et al*. (2002)	Low	Low	Medium	Medium low*
Ramsden (1992)	Medium	Medium	High	Medium high
Ramsden (1997)	High	Low	Low	Medium low*
Reid and Skryabina (2002)	Medium high	Medium	High	Medium high
Smith and Matthews (2000)	Medium low	Medium	Medium	Medium
Wierstra (1984)	Medium	Medium	Medium high	Medium
Wierstra and Wubbels (1994)	High	Low	Medium low	Medium low*
Yager and Weld (1999)	High	High	High	High
Zoller *et al*. (1990)	Medium low	Medium	Medium	Medium

* Study not designed to evaluate the effects of gender

Table 4.16: Weights of evidence for gender and understanding (GA)

Study	Weight of evidence A	Weight of evidence B	Weight of evidence C	Weight of evidence D
Smith and Bitner (1993)	Medium low	Medium high	Medium high	Medium
Wierstra (1984)	Medium	Medium	Medium high	Medium
Wierstra and Wubbels (1994)	High	Low	Medium low	Medium low*
Yager and Weld (1999)	High	High	High	High

* Study not designed to evaluate the effects of gender

Table 4.17: Weights of evidence for lower ability and attitude (LA)

Study	Weight of evidence A	Weight of evidence B	Weight of evidence C	Weight of evidence D
Huppert *et al*. (1992)	Low	Low	Medium	Medium low
Nentwig *et al*. (2002)	Low	Low	Medium	Medium low*
Smith and Matthews (2000)	Medium low	Medium low	Medium	Medium low
Yager and Weld (1999)	High	High	Medium	Medium high

* Study not designed to evaluate the effects of gender

Table 4.18: Weights of evidence for lower ability and understanding (LU)

Study	Weight of evidence A	Weight of evidence B	Weight of evidence C	Weight of evidence D
Huppert *et al*. (1992)	Low	Low	Medium	Medium low
Sutman and Bruce (1992)	Medium low	Medium low	Medium high	Medium
Yager and Weld (1999)	High	High	Medium high	High

4.4 Synthesis of the evidence

The synthesis consolidating evidence from this review draws primarily on the findings from studies weighted as *high*, *medium high* and *medium*. Findings from studies weighted as *medium low* are only considered if these corroborate findings of studies with a higher weight of evidence. Findings of other studies in the map may contribute to the synthesis as supporting evidence only. For instance, findings on the impact of short context-based interventions (as opposed to full context-based courses) relating to gender or lower-ability pupils will be used if they corroborate the synthesised findings of the in-depth review.

Only a small number of studies are considered in the in-depth review, and these are of variable quality. Therefore many of the review findings have, on purpose, been cast in tentative terms on account of their limited evidence base. For that reason the findings below have been reported under two headings: those supported by *reasonable evidence* and others supported by *some evidence*. Conclusions providing reasonable evidence are those that are based on converging findings of at least two studies with medium-to-high weight of evidence. No findings are claimed as being based on *strong evidence*.

4.4.1 The effect of context-based/STS approaches on boys and girls

The review suggests that there is ***reasonable*** evidence of the following:

a. *Girls in classes using a context-based/STS approach held significantly more positive attitudes to science than their female peers in classes using a traditional approach.*

   This evidence comes from three medium-to-high rated studies (Yager and Weld, 1999 – high; Reid and Skryabina, 2002 – medium high; Smith and Matthews, 2000 – medium). Attitudes to science in the study by Yager and Weld (1999) incorporate pupils' attitudes to the study of science, science classes and science teachers. The authors provide mean scores for pre-intervention and post-intervention responses for the experimental group (using a context-based approach) and the control group (using a traditional approach) [girls N=2,557 experimental group, pre-mean 40.7, post-mean 59.7, girls N=632 control group, pre-mean 46.9, post-mean 47.9, ANCOVA F1/f1=32.1 $p<0.01$]. Reid and Skryabina (2002) measure attitude to physics as pupils' enjoyment of school physics, perceived relevance of school physics and their self-perception in the study of physics. They conclude that girls following a context-based course in S4 hold more positive attitudes to physics than their (older) peers in S5/S6 who followed a traditional course [girls N=65 experimental (S4) group, girls N=28 control (S5/S6) group: for enjoyment χ^2

value =9.59, p<0.01, for usefulness χ^2 value =9.81, p< 0.01]. The study by Smith and Matthews (2000) probes the same aspects of attitudes to science. Supporting the same conclusion, they provide descriptive evidence only.

b. *Similarly, boys in classes using a context-based/STS approach held significantly more positive attitudes to science than their male peers in classes using a traditional approach.*

This evidence comes from three medium-to-high rated studies (Yager and Weld, 1999 – high; Reid and Skryabina, 2002 – medium high; Smith and Matthews, 2000 – medium). As indicated above, attitudes to science in the studies by Yager and Weld (1999) and Smith and Matthews (2000) incorporate pupils' attitudes to the study of science, science classes and science teachers. The large-scale pre-post experimental study by Yager and Weld (1999) provides solid evidence [boys N=2,673 experimental group: pre-test mean 47.4, post-test mean 61.5, N=688 control group: pre-test mean 47.8, post-test mean 52.8, ANCOVA F2/f2=22.3 p<0.01]. Again, Smith and Matthews (2000) only provide descriptive evidence. In their cross-sectional study, Reid and Skryabina (2002) concur that boys in S4 using a context-based approach hold significantly more positive attitudes to physics than their (older) peers in S5/S6 using a traditional approach (boys N=87 experimental (S4) group, boys N=68 control (S5/S6) group: for enjoyment χ^2 value =13.53, p<0.01, for usefulness χ^2 value =30.27, p< 0.01).

c. *A context-based/STS approach to teaching science narrowed the gap between boys and girls in their attitude to science.*

The evidence is based on five medium-to-high rated studies (Yager and Weld, 1999 – high; Ramsden, 1992 – medium high; Reid and Skryabina, 2002 – medium high; Smith and Matthews, 2000 – medium; Wierstra, 1984 – medium) and supplemented by two other studies (Ebenezer and Zoller, 1993 – medium low; Nentwig *et al.*, 2002 – medium low).

Yager and Weld (1999) conclude that girls in classes using a context-based/STS approach show significantly greater improvement in attitude to science than their male peers in the same classes [girls N=2,557 pre-mean 40.7, post-mean 59.7, boys N=2,673 pre-mean 47.4, post-mean 61.5, ANCOVA F3/f3=11.9 p<0.01]. Smith and Matthews (2000) draw the same conclusion but do not provide specific evidence in support. A similarity of post-intervention attitudes to science of boys and girls in classes using a context-based/STS approach is confirmed by Wierstra (1984) reporting that boys score significantly higher than girls on only one out of four aspects of attitude to physics, i.e. enjoyment of school physics [girls N=65, boys N=189: MANOVA p< 0.029. No detailed data]. Reid and Skryabina (2002) provide even stronger evidence of similarity across gender of a larger range of attitudes to science. Ramsden (1992) only reports post-intervention data. She uses enjoyment of contextualised science materials, practical work and non-practical work, perceived relevance to everyday events and interest in science as indicators for attitude to science. She found that there were few significant differences between overall attitudes of boys and girls in classes using a context-based approach [girls N=59, boys N=65, χ^2 analysis, no values provided]. Ebenezer and Zoller (1993), and Nentwig *et al.* (2002) also report finding no gender differences in attitude between boys and girls in context-based/STS courses, but provide no supporting data.

d. *In cases when boys enjoyed the materials significantly more than girls, this was due to the nature of the practical work in the unit. In cases when girls enjoyed context-based materials significantly more than boys, this was because of the non-practical activities in the unit.*

The evidence is drawn from three medium-to-high rated studies (Ramsden, 1992 – medium high; Reid and Skryabina, 2002 – medium high; Wierstra, 1984 – medium).

Ramsden (1992) finds larger male enjoyment of one (out of ten) context-based teaching unit mainly due to the difference in appreciation of the practical work included [girls N=59 mean=21.2, boys N=65 mean=25.4, χ^2 analysis, $p<0.01$]. Ramsden (1992) finds larger female enjoyment of one (out of ten) context-based teaching unit mainly due to the difference in appreciation of non-practical work, often seen as writing tasks [girls N=59 mean=22.5, boys N=65 mean=18.6, χ^2 analysis, $p<0.05$]. Reid and Skryabina (2002) find that the largest proportion of boys and girls in classes using a context-based approach consider 'doing practical work' their favourite activity, but this involves 88% of the boys, and only 52% of the girls. The opposite trend is apparent for the activity 'explaining natural phenomena' which is selected as a favourite activity by 33% of the boys and 46% of the girls [no significance levels calculated].

These conclusions are supported indirectly by the findings of Wierstra (1984). His data show that boys enjoy a context-based approach more than girls, and that enjoyment of physics correlates with the perceived level of inquiry (centred around notions of participation, problem-solving, investigative practical work) involved in the lessons. Thus, the greater amount of inquiry learning in the context-based approach in PLON has a more favourable affective outcome for boys (Wierstra, 1984).

The review suggests that there is ***some*** evidence of the following:

e. *Boys and girls in classes using a context-based approach significantly more often perceived a close link between science, technology and society than their gender peers in traditional classes. There were slight gender-related differences in the way science is linked to technology and society.*

This evidence is based on the study by Zoller *et al.* (1990 – medium). They used as measures for attitude to science the preferred locus of decision making on science-related world problems between the public and experts; the perceived responsibility of scientists for the applications of their discoveries; the influence of values on interpreting scientific facts; and the role of science and technology. Zoller *et al.* (1990) report that boys in the experimental group favoured the public (against scientists and engineers) deciding on issues of world food production [boys N=41 experimental group, boys N=130 control group: χ^2 value =9.00 $p<0.025$]. Also, these boys significantly more often than their male peers in traditional classes believed that personal opinion and moral values play a role in interpretation of facts, not that scientists are neutral and objective. [boys N=41 experimental group, boys N=130 control group: χ^2 value =16.35, $p<0.005$].

On the other hand, Zoller *et al.* (1990) report that girls in the experimental group favoured the public (against scientists and engineers) deciding on issues of world food production [girls N=60 experimental group, girls N=146 control group: χ^2 value =8.38 $p<0.025$]. Also, these girls are reported significantly more often than their peers in traditional classes considering scientists to be responsible for the consequences of their discoveries, not only for the quality of the science *per se* [girls N=60 experimental group, girls N=146 control group: χ^2 value =16.22, p<0.005].

f. *Girls in classes using a context-based/STS approach developed a significant more positive attitude towards taking a science career as compared to boys in these classes.*

The conflicting evidence for this conclusion is drawn from two medium-to-high rated studies (Yager and Weld, 1999 – high; Reid and Skryabina, 2002 – medium high). Yager and Weld (1999) use a specific instrument to gather data from Grade 6-8 pupils about their career intentions [girls N=187, boys N=177, ANCOVA of post-test responses to four out of five items (with pre-test scores as covariate) show difference of p<0.05 level. F values between 97.05 and 15.06]. No difference emerged in career perceptions between boys and girls in classes using a traditional approach**.** In contrast, Reid and Skryabina (2002) find that the same, and very high, proportions of boys and girls at S4 level (using a context-based approach) suggest they wish to continue with physics courses at higher levels (92% and 89% respectively).

g. *Boys and girls in classes using a context-based/STS approach showed significantly better conceptual understanding of science than their gender peers in classes using a traditional approach.*

This evidence comes from Yager and Weld (1999 – high). Conceptual understanding in the study by Yager and Weld incorporates pupils' understanding of science concepts and the application of these to new situations. They provide mean scores for pre-intervention and post-intervention responses for the experimental group (using a context-based approach) and the control group (using a traditional approach). For each of conceptual understanding and application, the evidence is provided for boys [for concepts boys N=2,673 experimental group, pre-mean=35.8, post-mean=73.7, N=688 control group, pre-mean=36.1, post-mean=52.1, T-test F2/f2=22.7 p<0.01; for applications boys N=2,673 experimental group, pre-mean=34.2, post-mean=74.7, N=688 control group, pre-mean=33.9, post-mean=50.7, T-test F2/f2=20.8 p<0.01] and for girls [for concepts girls N=2,557 experimental group, pre-mean=32.6, post-mean=71.8, n=632 control group, pre-mean=32.8, post-mean=47.5, T-test F1/f1=23.4 p<0.01; for applications girls N=2557 experimental group, pre-mean=29.7, post-mean=74.2, N=632 control group, pre-mean=30.0, post-mean=44.2, T-test F1/f1=26.3 p<0.01].

h. *Girls in classes using a context-based/STS approach showed equal conceptual understanding of science as male peers in the same classes.*

This conclusion is contentious. The evidence is based on two medium-to-high rated studies (Yager and Weld, 1999 – high; Wierstra, 1984 – medium). The findings of Yager and Weld support this conclusion fully [concept girls N=2,557 pre-mean=32.6, post-mean=71.8, boys N=2,673 pre-mean=35.8, post-mean=73.7, T-test F3/f3=3.7, no difference; application girls N=2,557 pre-mean=29.7, post-mean=74.2, boys N=2,673 pre-mean=34.2, post-mean=74.7, T-test F3/f3=4.8, no difference]. In contrast, Wierstra (1984) reports that boys in classes using a context-based approach score significantly higher than girls on both achievement tests [girls N=65, boys N=189: MANOVA p< 0.040 and p<0.0001 respectively (no detailed data).

4.4.2 The effect of context-based/STS approaches on lower-ability pupils

The review suggests that there is ***some*** evidence of the following:

a. *Lower-ability pupils in classes using a context-based/STS approach held significantly more positive attitudes to science than lower-ability pupils in classes using a traditional approach.*

This evidence comes from Yager and Weld (1999 – medium high) and is supported by a subsidiary study (Smith and Matthews, 2000 – medium low). The study by Yager and Weld defines lower-ability pupils as the bottom-achieving third within their mixed ability sample. Attitudes to science in the same study incorporate pupils' attitudes to the study of science, science classes and science teachers. They provide mean scores for pre-intervention and post-intervention responses for lower-ability pupils in the experimental group (using a context-based approach) and the control group (using a traditional approach) [lower ability N=1,754 experimental group, pre-mean=32.8, post-mean=52.1, n=438 control group, pre-mean=32.1, post-mean 40.1, ANCOVA F1/f1=46.7 $p<0.01$]. The study by Smith and Matthews (2000) probes the same aspects of attitudes to science. They provide descriptive evidence only [N=37 experimental group, significant positive change in post-test responses to 19/25 items, N=23 control group, significant positive change in post-test responses to 7/25 items, significant negative change in post-test responses to 1/25 item].

b. *Lower-ability pupils in classes using a context-based/STS approach developed significantly more positive attitudes towards science than high ability peers in the same classes.*

The evidence is based on the study by Yager and Weld (1999 – high) with support from a subsidiary study (Nentwig *et al.*, 2002 – medium low). They compare within the experimental group pre- and post-intervention attitude scores of the top (higher ability) and bottom (lower ability) third of their large mixed ability sample [lower ability N=1,754 pre-mean=32.8, post-mean=52.1, higher ability N=1,757 pre-mean=51.8, post-mean=75.7, ANCOVA F3/f3=34.5 $p<0.01$]. Nentwig *et al.* (2002) provide only descriptive results and concur that pupils with lower chemistry grades in the previous school year showed a higher increase in interest in chemistry lessons, their perceived relevance and instructional quality.

c. *Lower-ability pupils in classes using a context-based/STS approach showed significant better conceptual understanding of science than their lower ability peers in classes using a traditional approach.*

This evidence comes from two medium-to-high rated studies (Yager and Weld, 1999 – high; Sutman and Bruce, 1992 – medium). Both studies divide conceptual understanding into pupils' understanding of science/chemistry concepts and the application of these concepts to new situations. Yager and Weld (1999) provide mean scores for pre-intervention and post-intervention responses for lower-ability pupils in the experimental group (using a context-based approach) and the control group (using a traditional approach) [for concepts lower ability N=1,754 experimental group, pre-mean=20.4, post-mean=56.1, N=438 control group, pre-mean=21.4, post-mean=29.1, T-test F1/f1=33.1 $p<0.01$; for application lower ability N=1,754 experimental group, pre-mean=19.7, post-mean=59.3, N=438 control group, pre-mean=20.3, post-mean=29.8, T-test F1/f1=36.3 p<0.01]. Sutman and Bruce (1992) provide concurring descriptive results but without detailed data from their large study of 3,700 pupils divided over experimental and control groups. Both groups consist of non-science majors with generally below average achievement.

d. *Lower-ability pupils in classes using a context-based/STS approach showed higher gain in conceptual understanding of science than high ability peers in the same classes.*

   The evidence is based on the study by Yager and Weld (1999 – high). Within the experimental group, the authors report pupils' understanding of science concepts and the application of these concepts to new situations for the top and bottom third of the sample. [for concepts lower ability N=1,754 experimental group, pre-mean=20.4, post-mean=56.1, higher ability N=1,757, pre-mean=42.1, post-mean=80.2, T-test F3/f3=16.7 p<0.01; for application lower ability N=1,754 experimental group, pre-mean=19.7, post-mean=59.3, higher ability N=1,757, pre-mean=40.1, post-mean=81.2, T-test F3/f3=18.4 p<0.01].

Links with other reviews

No other reviews of the learning effects of context-based/STS approaches to science teaching for boys, girls or lower-ability pupils have been undertaken. It was therefore not possible to compare the findings of this review with those of other reviews.

4.5 In-depth review: quality-assurance results

The quality-assurance processes for in-depth reviewing described in section 2.3.5 were followed. No areas of significant disagreement remained after moderating the data-extraction summaries between the pairs of experts. Generally, guidelines by collaborators from the EPPI-Centre were followed. The algorithm for determining the weighting of categories B and C (Appendix 2.4) worked well in securing coherence of these judgements across data-extraction teams. Additionally, three members of the Review Group independently ranked the studies they data-extracted on the basis of what they felt was the overall quality. Rankings were consistent and allowed for the construction of an overall ranking.

5. FINDINGS AND IMPLICATIONS

5.1 Summary of principal findings

5.1.1 Identification of studies

As in the 'base' review (Bennett *et al*., 2003), the overall research question for this review is: ***What is the effect on pupils of teaching approaches which emphasise placing science in context and promote links between science, technology and society (STS)?***

Within this, two in-depth review questions were identified:

What is the evidence from evaluative studies of the effect of context-based or STS courses on the attitude to science and/or the understanding of science ideas of boys and girls in the 11 to 16 age range?

What is the evidence from evaluative studies of the effect of context-based or STS courses on the attitude to science and/or on the understanding of science ideas of lower-ability pupils in the 11 to 16 age range?

5.1.2 Mapping of all included studies

Sixty-one studies met the inclusion criteria developed for the overall research review. These studies were keyworded and formed the basis of the systematic map. The map revealed a number of characteristics of research on the effect of context-based and STS approaches to teaching science, as summarised below.

- One in three of the studies report work that has taken place in the USA, and one in four report work in the UK.
- Past and currently active research groups in this area are based at the Universities of British Columbia (Canada), Cambridge (UK), Iowa (USA), Utrecht (The Netherlands) and York (UK).
- Two in three studies were undertaken with 11- to 16-year-old pupils, the remainder with pupils between the ages of 17 and 20.
- More than half the studies report on context-based/STS interventions in science, the remainder being equally split between chemistry and physics. Very little research has been done on context-based/STS approaches in the teaching of biology.
- Almost two out of three studies evaluated interventions using context-based approaches, the majority of which concern full courses. The remaining third evaluated interventions using STS approaches equally divided between STS-enrichment activities and full STS courses.
- About 60% of the studies are naturally occurring evaluations, and the remaining 40% are researcher-manipulated.

- The most popular techniques for gathering data are test results, questionnaires, (dis)agreement scales and interviews. The first is over-represented amongst researcher-manipulated evaluations, the last amongst naturally occurring evaluations.

5.1.3 Nature of studies selected for in-depth review

Fourteen studies met the inclusion criteria for the in-depth review. Table 5.1 summarises the overall weights of evidence assigned to each of these studies.

Table 5.1: Overall weights of evidence assigned to studies

		Evaluation aspect			
		Gender		**Lower ability**	
		Attitude (GA)	**Understanding (GU)**	**Attitude (LA)**	**Understanding (LU)**
Ebenezer and Zoller (1993)		ML*			
Greenlee and Lambert (1996)		ML			
Huppert *et al.* (1992)				ML	ML
Nentwig *et al.* (2002)		ML*		ML*	
Ramsden (1992)		MH			
Ramsden (1997)		ML*			
Reid and Skryabina (2002)		MH			
Smith and Bitner (1993)			M		
Smith and Matthews (2000)		M		ML	
Sutman and Bruce (1992)					M
Wierstra (1984)		M	M		
Wierstra and Wubbels (1994)		ML*	ML*		
Yager and Weld (1999)		H	H	MH	H
Zoller *et al.* (1990)		M			
Total frequencies	*H*	1	1	0	1
	MH	2	0	1	0
	M	3	2	0	1
	ML	5	1	3	1
	L				
	Total	***11***	***4***	***4***	***3***

L = Low, ML = Medium low, M = Medium, MH = Medium high, H = High
* Study not designed to evaluate effects of gender or lower ability

5.1.4 Synthesis of findings from studies in the in-depth review

The review suggests that there is ***reasonable*** evidence of the following:

- *Girls in classes using a context-based/STS approach held significantly more positive attitudes to science than their female peers in classes using a traditional approach* (based on mutually supportive evidence from one study rated high, one rated medium high, and one study rated medium).

- *Similarly, boys in classes using a context-based/STS approach held significantly more positive attitudes to science than their male peers in classes using a traditional approach* (based on mutually supportive evidence from the same set of studies as above).

- *A context-based/STS approach to teaching science narrowed the gap between boys and girls in their attitude to science* (based on mutually supportive evidence from one study rated high, two rated medium-high, and two studies rated medium; supplementary support from two other studies).

- *In cases when boys enjoyed the materials significantly more than girls, this was due to the nature of the practical work in the unit. In cases when girls enjoyed context-based materials significantly more than boys, this was because of the non-practical activities in the unit* (based on mutually supportive evidence from two studies rated medium high, and one study rated medium).

The review suggests there is ***some*** evidence of the following:

- *Boys and girls in classes using a context-based approach significantly more often perceived a close link between science, technology and society than their gender peers in traditional classes. There were slight gender-related differences in the way science was linked to technology and society* (based on evidence from one study rated medium).

- *Boys and girls in classes using a context-based/STS approach showed significantly better conceptual understanding of science than their gender peers in classes using a traditional approach* (based on evidence from one study rated high).

- *Girls in classes using a context-based/STS approach developed a significantly more positive attitude towards taking a science career as compared with boys in these classes* (based on contradicting evidence from one study rated high, supporting this conclusion, and one study rated medium high, concluding that girls and boys have the same, both very positive, attitude to pursuing a science career).

- *Girls in classes using a context-based/STS approach showed equal conceptual understanding of science as male peers in the same classes* (based on contradictory evidence from one study rated high, supporting the conclusion, and one study rated medium, concluding that boys using a context-based/STS approach significantly outperform girls using the same approach).

- *Lower-ability pupils in classes using a context-based/STS approach held significantly more positive attitudes to science than lower-ability pupils in classes using a traditional approach* (based on evidence from one study rated high; supplementary support from one other study).

- *Lower-ability pupils in classes using a context-based/STS approach developed significantly more positive attitudes towards science than high-ability peers in the same classes* (based on evidence from one study rated high; supplementary support from one other study).

- *Lower-ability pupils in classes using a context-based/STS approach showed significantly better conceptual understanding of science than their lower-ability peers in classes using a traditional approach* (based on mutually supportive evidence from one study rated high and one study rated medium).

- *Lower-ability pupils in classes using a context-based/STS approach showed higher gain in conceptual understanding of science than high-ability peers in the same classes* (based on evidence from one study rated high).

5.2 Strengths and limitations of this systematic review

Strengths

- The review focus is highly topical. The current concern about re-engaging male learners at secondary school level in their own schooling has reignited interest in gender issues. Further evidence of the topicality comes from the range of countries in which studies have been undertaken.

- The review has established that there is consistency in the research approaches adopted by those researching the effect of context-based or STS approaches to science teaching on pupils' understanding of science ideas or on their attitude to science. Such approaches make use of a pre-post quasi-experimental research design and generate quantitative data. For measuring understanding of science ideas, standardised achievement tests are used to compare the effect of different curricula. For measuring attitude to science, a large variety of instruments using Likert- or Osgood-type scales are used.

- Quality-assurance results were high for all stages of the review.

Limitations

The review has two main limitations:

- There was a scarcity of studies that focused on the effects of context-based or STS approaches for boys and girls, or for lower-ability pupils, as an independent variable. This resulted in very little evidence related specifically to the in-depth review questions, especially to the question of lower ability. Of these studies, only about half use a comparison group for the teaching approach. Only seven studies were judged to be of reasonable quality with respect to the review question: that is, they had an overall weight of evidence of medium or higher.

- Although the studies in the in-depth review shared a number of similar characteristics at the broad level, there were considerable differences at the detailed level. For example, a considerable range of aspects of attitude to science was being measured, and the measurement of understanding of science ideas also varied widely according to the topic that happened to have been taught during the intervention. Thus, a number of the findings appeared to be very specific to the particular study from which they emerged rather than suggestive of any overall patterns. The synthesis could only claim *reasonable* or *some* evidence (as opposed to *strong* evidence) for any of the conclusions drawn.

Additionally, the Review Group feel some concern about the number of studies which had to be included in the in-depth review based on the fact that they were evaluations, and mentioned gender or lower ability, but did not evaluate these central aspects of the review. However, this problem is a function of the review process itself, rather than this specific review.

Resource limitations have resulted in the omission of publications in languages other than English and of dissertations from outside the UK. This may have limited the depth of the review.

5.3 Implications

The Review Group is cautious about commenting on implications of the review for policy and practice for the reasons given in the preceding section on 'Limitations'.

5.3.1 Policy

The review has yielded reasonable evidence that both girls and boys in secondary school science classes using a context-based or STS approach develop more positive attitudes that peers following more traditional courses. The review also supports the conclusion that the differences between the attitudes to science of boys and girls can be decreased by learning through context-based or STS approaches.

The review therefore indicates that a policy which embraces the teaching of science through linking it to the everyday experiences of pupils is likely to have a beneficial effect on their attitudes to science. However, it should also be noted that there is a scarcity of high quality research evidence of the positive effects of such an approach for different subsets of pupils, such as lower-ability learners (as indicated by this review), upper-ability learners, those from cultural minorities, second-language speakers, or those using different learning styles or sense-making strategies.

The review provides some evidence that context-based/STS approaches may foster more positive attitudes to science careers in girls in particular, although the data gathered relate only to indications of intent, not actual choices. However, there is no evidence to suggest that a policy of using context-based/STS approaches has a detrimental effect on attitudes to science careers of boys or girls.

5.3.2 Initial teacher training

It is highly desirable that students and tutors on initial teacher training courses should be made aware of the evidence on the effects on context-based and STS approaches on pupils' attitudes, as should those who produce resources for use on such courses. This includes the DfES (which is responsible for introducing the Key Stage 3 Strategy), QCA (which has produced an extensive and detailed Scheme of Work for Key Stage 3), and the TTA (which specifies standards for newly qualified teachers).

It is clear from this review that context-based/STS approaches stimulate pupils' interest in science. However, there is a possible tension between using contexts as starting points and the current emphasis in many of the materials produced for science teaching which indicate lessons should start by making learning outcomes explicit to pupils. The two notions are not incompatible, but require a more sophisticated approach to lesson planning. A strategy of explicitly stating the intended lesson outcomes after a context-based lesson introduction may well avoid one of the main risks of context-based learning: that is, some pupils' idea that the learning outcome is about the understanding of the context rather than the science concepts underlying the context.

5.3.3 Practice

In order to improve attitudes to science of boys and girls, especially interest in science lessons, teachers should use everyday interests as starting points. Several of the studies in the map provide ideas for possible contexts and teaching strategies: for instance, see Kortland (1997), Ramsden (1992), Ruba *et al.* (1991) and Solomon (1992).

In order to gain maximum benefit from the use of a context-based/STS approach for both girls and boys, teachers should consider making use of a variety of activities in their lessons, though should they adopt context-based or STS programmes this is going to happen as a matter of course.

5.3.4 Research

Secondary research

One particular area of the systematic map would benefit from further exploration. This relates to studies which have reported on the effects on understanding of science ideas. Present knowledge in this area synthesised through previous systematic reviews is seriously limited. Currently, only those studies which have explored understanding and attitude, or understanding in relation to gender or ability, have formed the basis of in-depth reviews. A broadening of the evidence-base using the existing map represents an efficient use of resources.

In the light of the effectiveness of context-based/STS approaches in improving pupils attitudes to science, and the recommendation for adoption of such approaches in classrooms and ITT programmes, a consolidation of knowledge is required about teacher professional development strategies used for supporting a change practice: which characteristics of professional development strategies lead to adoption (or adaptation) of the use of context-based approaches?

What are the crucial steps in the development of teacher ownership of context-based/STS approaches?

Primary research

This review points to three specific areas which would benefit from primary research. Firstly, the review has indicated a dearth of work on the effects of context-based/STS approaches on lower-ability pupils, with much of the evidence emerging from one study. Thus this group would appear to be one for whom additional research would be useful. Secondly, the review has also suggested that little research has been done into the effects on more able pupils, again pointing to an area of further work. Thirdly, given the current concerns about boys' perceived underachievement and the indication in this review (based on limited evidence and from the USA only) that context-based/STS approaches enhanced boys' understanding of science, it would be desirable to explore aspects of gender and achievement for context-based and STS courses in more detail.

More generally, both this review, and the attitude and understanding review (Bennett *et al.*, 2003) have tended to treat quite sophisticated concepts as fairly simple ideas. Some of these would benefit from further unpicking. For example, the term ‘context’ can be interpreted in a wide variety of ways, and it would be useful to look at some of these ways, with a view to establishing how particular groups of pupils respond to different contexts. Such work would be particularly useful for those developing courses on public understanding of science.

By focusing on evaluations involving an experimental design, both reviews have also tended to yield evidence which relates to *what* effects context-based/STS approaches have had, rather than *why* these effects have occurred. Thus some empirical work involving interviews with pupils to explore their views on particular contexts and their particular effects would be useful. Such work could also be fruitful in exploring, for example, aspects such as why more positive attitudes to science do not appear to be translated to any significant extent into a desire to pursue careers in science.

6. REFERENCES

6.1 Studies included in the systematic map

The 61 studies included in the systematic map were reported in 67 papers. For the purpose of the map and synthesis, one paper was selected as the lead paper for each study. Subsidiary papers are marked with an asterisk.*

Alsop S, Watts M, Hanson J (1998) Pupils' perceptions of radiation and radioactivity: the wary meet the unsavoury. *School Science Review* **79**: 75–79.

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

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

Barker V, Millar R (1996) *Differences Between Salters' and Traditional A-level Chemistry Students' Understanding of Basic Chemical Ideas*. York, UK: University of York.

*Barker V, Millar R (1999) Students' reasoning about chemical reactions: what changes occur during a context-based post-16 chemistry course? *International Journal of Science Education* **21:** 645–665.

Barker V, Millar R (2000) Students' reasoning about basic chemical thermodynamics and chemical bonding: what changes occur during a context-based post-16 chemistry course? *International Journal of Science Education* **22:** 1171–1200.

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

Campbell B, Hogarth S, Lubben F (2000a) Contextualising the physics curriculum: learners' perceptions of interest and helpfulness. Paper presented at the Annual Meeting of the British Educational Research Association. Cardiff, UK: 7–9 September.

Campbell B, Lubben F, Dlamini Z (2000b) Learning science through contexts: helping pupils make sense of everyday situations. *International Journal of Science Education* **22:** 239–252.

Christie DJ, Nelson L (1988) Student reactions to nuclear education. *Bulletin of the Atomic Scientists* **44:** 22–23.

Dahncke H, Behrendt H, Reiska P (2001) A comparison of STS-teaching and traditional physics lessons on the correlation of physics knowledge and taking action. In: Dahncke H, Behrendt H (eds) *Research in Science Education – Past, Present and Future*. Netherlands: Kluwer Academic Publishers, pages 77–82.

Dlamini B, Lubben F, Campbell B (1996) Liked and disliked learning activities: responses of Swazi students to science materials with a technological approach. *Research in Science and Technological Education* **14:** 221–236.

Ebenezer JV, Zoller U (1993) Grade 10 students' perceptions of and attitudes toward science teaching and school science. *Journal of Research in Science Teaching* **30:** 175–86.

Eijkelhof H, Lijnse P (1988) The role of research and development to improve STS education: experiences from the PLON project. *International Journal of Science Education* **10:** 464–474.

Ferreira M (2001) Building communities through role models, mentors, and hands-on-science. *School Community Journal* **11:** 27–37.

Fowler C, Sinclair A (1992) Forecasting the future: a student vision quest. *Science Teacher* **59:** 50–53.

Gallagher SA, Stepien WJ, Rosenthal H (1992) The effects of problem-based learning on problem-solving. *Gifted Child Quarterly* **36:** 195–200.

Greenlee S, Lambert L (1996) *The Science of Living Spaces: Women in the Environment of the 21st Century. An Evaluation and Guide Book.* Virginia, USA: Christopher Newport University.

Helms JV (1998) Science and/in the community: context and goals in practical work. *International Journal of Science Education* **20:** 643–653.

Hughes G (2000a) Marginalization of socioscientific material in science-technology-society science curricula: some implications for gender inclusivity and curriculum reform. *Journal of Research in Science Teaching* **37:** 426–440.

*Hughes G (2000b) Salters' curriculum projects and gender inclusivity in science. *School Science Review* **81:** 85–89.

Huppert J, Simchoni D, Lazarowitz R (1992) Human health and science: a model for an STS high school biology course. *American Biology Teacher* **54:** 395–400.

Key M-B (1998) *Students' Perceptions of Chemical Industry; Influences of Course Syllabi, Teachers, Firsthand Experience*. York, UK: University of York.

Kortland J (1997) Garbage: dumping, burning and reusing/recycling: students' perception of the waste issue. *International Journal of Science Education* **19:** 65–77.

Kortland J (2001) A problem-posing approach to teaching for scientific literacy: the case of decision-making about packaging waste. In: De Jong O, Savelsbergh JR, Alblas A (eds) *Teaching for Scientific Literacy*. Utrecht, Netherlands: CDB Press, pages 87–98.

*Lenton GM (1990) SATIS 16–19: a preliminary review of school trials. *School Science Review* **71:** 135–140.

Lenton GM (1991) A review of SATIS 16-19 national trials. *School Science Review* **73:** 7–18.

Licht P (1990) Teaching physics and context in the Netherlands. In: Krabbedam H, De Vries M (eds) *Goals and Methods in Science, Mathematics and Technology Education*. Enschede, Netherlands: National Institute for Curriculum Development, pages 65–78.

Lubben F, Campbell B, Dlamini B (1996) Contextualizing science teaching in Swaziland: some student reactions. *International Journal of Science Education* **18:** 311–320.

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:** 26–40.

Monhardt RM, Monhardt L (1998) Facilitating science literacy in a rural school. *Bulletin of Science, Technology and Society* **18:** 47–53.

Nentwig P, Parchmann I, Demuth R, Graesel C, Ralle B (2002) Chemie im Kontext: from situated learning in relevant contexts to a systematic development of basic chemical concepts. Paper presented at the Symposium on Context-Based Science Curricula. Kiel, Germany: October 10–13.

Pedersen JE (1992) The effects of a cooperative controversy, presented as an STS issue, on achievement and anxiety in secondary science. *School Science and Mathematics* **92:** 374–380.

Phillips LM, Norris SP (1999) Interpreting popular reports of science: what happens when the reader's world meets the world on paper? *International Journal of Science Education* **21:** 317–27.

Pilling G, Holman J, Waddington D (2001) The Salters' experience. *Education in Chemistry* **38:** 131–133.

Ramsden JM (1992) If it's enjoyable, is it science? *School Science Review* **73:** 65–71.

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

Rannikmae M (2001) Effectiveness of teacher-developed scientific and technological literacy materials. In: De Jong O, Savelsbergh JR, Alblas A (eds) *Teaching for Scientific Literacy*. Utrecht, Netherlands: CDB Press, pages 71–86.

Reid N, Skryabina E (2002) Attitudes towards physics. *Research in Science and Technological Education* **20:** 67–81.

Roth WM (2001) Learning science through technological design. *Journal of Research in Science Teaching* **38:** 768–790.

Rubba PA (1990) STS education in action: what researchers say to teachers. *Social Education* **54:** 201–203.

Rubba PA, McGuyer M, Wahlund TM (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:** 537–552.

*Skryabina E (2000) Students' attitudes to learning physics at school and university levels in Scotland. University of Glasgow, UK. Unpublished PhD thesis.

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

Smith LA, Bitner BL (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. Kansas City, MO, USA: April 1–4.

Solomon J (1988) Science technology and society courses: tools for thinking about social issues. *International Journal of Science Education* **10:** 379–387.

Solomon J (1989) Discussing nuclear power. *Physics Education* **24:** 344–347.

Solomon J (1992) The classroom discussion of science-based social issues presented on television: knowledge, attitudes and values. *International Journal of Science Education* **14:** 431–444.

Stevens LP (2001) 'South Park' and society: instructional and curricular implications of popular culture in the classroom. *Journal of Adolescent and Adult Literacy* **44:** 548–555.

Sutman F, Bruce M (1992) Chemistry in the community – ChemCom. *Journal of Chemical Education* **69:** 564–567.

Truex RT (1987) *Township of Ocean School District Contemporary Science. Program description.* New Jersey, USA: Ocean Township Board of Education.

*Tsai C-C (1999) The progression toward constructivist epistemological views of science: a case study of the STS instruction of Taiwanese high school female students. *International Journal of Science Education* **21:** 1201–1222.

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:** 1099–1115.

Van Weelie D (2001) Contextualising biodiversity. In: De Jong O, Savelsbergh JR, Alblas A (eds) *Teaching for Scientific Literacy*. Utrecht, Netherlands: CDB Press, pages 99–116.

Walker D (1990) The evaluation of SATIS. *School Science Review* **72:** 31–39.

Wierstra RFA (1984) A study on classroom environment and on cognitive and affective outcomes of the PLON-curriculum. *Studies in Educational Evaluation* **10:** 273–282.

Wierstra RFA, 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:** 437–455.

Wiesenmayer RL, Rubba PA (1999) The effects of STS issue investigation and action instruction versus traditional life science instruction on seventh grade students' citizenship behaviors. *Journal of Science Education and Technology* **8:** 137–144.

Wiley DA (1991) Observations on a science-technology-society course for non-science students. *Perceptual and Motor Skills* **72:** 845–846.

Williams RA, Lisowski M, Niemietz K (1990) *The Illinois Rivers Project.* Illinois, USA: Southern Illinois University of Edwardsville.

Winther AA, Volk TL (1994) Comparing achievement of inner-city high school students in traditional versus STS-based chemistry courses. *Journal of Chemical Education* **71:** 501–505.

Yager RE, Blunck SM, Binadja A, McCormick RW, Penick JE (1988) *Assessing Impact of S/T/S Instruction in 4–9 Science in Five Domains.* Iowa, USA: University of Iowa.

Yager RE, Tamir P (1993) STS approach: reasons, intentions, accomplishments, and outcomes. *Science and Education* **77:** 637–658.

Yager RE, Weld JD (1999) Scope, sequence and co-ordination: The Iowa Project, a national reform effort in the USA. *International Journal of Science Education* **21:** 169–194.

Zoller U, Ebenezer JV, Morley K, Paras S, Sandberg V, West C, Wolthers T, Tan SH (1988) *Goals' Attainment in Science-Technology-Society (S/T/S) Education: Expectations and Reality. A probe into the case of British Columbia.* British Columbia, Canada: Science Education Research.

Zoller U, Ebenezer JV, Morley K, Paras S, Sandberg V, West C, Wolthers T, Tan SH (1990) Goal attainment in science-technology-society (S/T/S) education and reality: the case of British Columbia. *Science Education* **74:** 19–36.

*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:** 25–36.

6.2 Other references used in the text of the report

Aikenhead G (1994) What is STS teaching? In: Solomon J, Aikenhead G (eds) *STS Education: International Perspectives On Reform.* New York: Teachers College Press, pages 47–59.

American Chemical Society (ACS) (1988) *ChemCom: Chemistry in the Community*. Dubuque, IA, USA: Kendall-Hunt.

Association for Science Education [ASE] (1978) *Science for Less Academically Motivated Pupils (the 'LAMP' project).* Hatfield: ASE.

Barker V, Millar R (2000) Students' reasoning about basic chemical thermodynamics and chemical bonding: what changes occur during a context-based post-16 chemistry course? *International Journal of Science Education* **22:** 1171–1200.

Bennett J (2003) *Teaching and Learning Science: A Guide to Recent Research and its Applications.* London: Continuum.

Bennett J, Holman J (2003) Context-based approaches to the teaching of chemistry: what are they and what are their effects? In: Gilbert J (ed.) *Chemical Education: Towards Research-Based Practice.* Dordrecht, Netherlands: Kluwer Academic Publishers, pages 165–184.

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, University of London. http://eppi.ioe.ac.uk/reel/

Bennett J, Lubben F, Hogarth S, Campbell B (2004) 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. http://eppi.ioe.ac.uk/reel/

Borgford C (1995) The Salters Science materials: a study of teachers' use and areas of focus. Unpublished DPhil thesis. York: University of York.

Burton W, Holman J, Pilling G, Waddington D (1994) *Salters Advanced Chemistry*. Oxford: Heinemann.

Campbell R, Lazonby J, Millar R, Nicolson P, Ramsden J, Waddington D (1994) Science: the Salters approach; a case study of the process of large scale curriculum development. *Science Education* **78:** 415–447.

Davies F, Greene T (1984) *Reading for Learning in the Sciences.* Edinburgh: Oliver Boyd.

Dutch Physics Curriculum Development Project (Projekt Leerpakketontwikkeling Natuurkunde [PLON]) (1988) Utrecht, Netherlands: Rijksuniversteit Utrecht, Vakgroep Natuurkunde-Didactiek.

Eijkelhof H, Kortland K (1988) Broadening the aims of physics education. In: Fensham P (ed.) *Development and Dilemmas in Science Education.* New York: Falmer Press, pages 282-385.

EPPI-Centre (2002a*) Core Keywording Strategy (Version 0.9.6).* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

EPPI-Centre (2002b) *EPPI-Reviewer (Version 0.9.7)*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

EPPI-Centre (2002c) *Guidelines for Extracting Data and Quality Assessing Primary Studies in Educational Research (Version 0.9.7).* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

House of Commons (2002) *Science Education from 14-19. Third Report of the Science and Technology Committee.* London: The Stationery Office.

Hunt A, Millar R (eds) (2000) *AS Science for Public Understanding*. Oxford: Heinemann Educational.

Kortland J (1997) Garbage: dumping, burning and reusing/recycling: students' perception of the waste issue. *International Journal of Science Education* **19:** 65–77.

Kyriacou C (1998) *Essential Teaching Skills.* 2nd edition. Cheltenham: Stanley Thornes.

Lewis J (1981) *Science in Society*. London: Heinemann Educational Books.

Millar R, Osborne J (eds) (1998*) Beyond 2000: Science Education for the Future*. London: King's College.

Newton P, Driver R, Osborne J (1999) The place of argumentation in the pedagogy of school science. *International Journal of Science Education* **21:** 553–576.

Nuffield Foundation (1980) *Nuffield Science 13-16.* York: Longman.

Osborne J, Erduran S, Simon S, Monk M (2001) Enhancing the quality of argument in school science. *School Science Review* **82:** 63–70.

Sjøberg S (2000) Interesting all children in 'science for all'. In: Millar R, Leach J, Osborne J (eds) *Improving Science Education: The Contribution of Research.* Buckingham: Open University Press, pages 165-186.

Solomon J (1992) The classroom discussion of science-based social issues presented on television: knowledge, attitudes and values. *International Journal of Science Education* **14:** 431–444.

Supported Learning in Physics Project (SLIP) (1997) *Eight Units for Advanced Level.* Buckingham: Open University Press.

Tal R, Dori Y, Keiny S, Zoller U (2001) Assessing conceptual change of teachers involved in STES education and curriculum development – the STEMS project approach. *International Journal of Science Education* **23:** 247–262.

University of York Science Education Group (UYSEG) (1990-1992) *Science: the Salters Approach: 22 Unit Guides for Key Stage 4.* York/Oxford: UYSEG/Heinemann Educational.

Wray D, Lewis M (1997) *Extending Literacy: Children Reading and Writing Non-Fiction.* London: Routledge.

Yager R, Casteel J (1968) The University of Iowa science and culture project. *School Science and Mathematics* **67:** 412–416.

Appendix 2.1: Inclusion and exclusion criteria

TTA systematic review: inclusion/exclusion criteria

Review area: the effect on students of teaching approaches which emphasise placing science in context or promoting links between science, technology and society (STS)

Exclusion criteria: to be included, a study must *not* fall into any one of the following categories

EXCLUSION ON SCOPE

1. Not reporting on learning/teaching of science
   - definition of science: one or several of the school subjects integrated/general science, science, biology, chemistry physics or earth science

2. Not about context-based or STS learning/teaching
   - definition of context-based: using learners' environment and experiences as starting points for teaching)
   - definition of STS: intending to promote links between science and technology or between science and society

3. Not about the effect on students' understanding or attitude

4. Not about learners aged 11 to 18, or main focus not on learners aged 11 to 18

EXCLUSION ON STUDY TYPE

5. (a) Editorials, commentaries, book reviews
   (b) Policy documents
   (c) Resources
   (d) Bibliography
   (e) Theoretical (non-empirical) paper
   (f) Methodology paper
   (g) Review papers

EXCLUSION ON SETTING IN WHICH STUDY WAS CARRIED OUT

6. Not published in English

7. Not published in the period 1980–2002

Appendix 2.2: Search strategy for electronic databases

Subject
Teaching approaches which emphasise placing science in context and promote links between science, technology and society

Population
Pupils aged 11 to 18

Limits
English language
1980 to date

1. Databases

1a. ERIC

ERIC was searched on 10 February 2003, using the BIDS Ovid interface. The database was searched for the period 1980–2002 and 1,928 records were retrieved.

1. *(context-based or context-led or application-based or application-led)*
2. (contextualis$ or contextualiz$ or technological approach)
3. (Salters or SATIS or PLON or ChemCom or Science in context)
4. *(science-technology-society or (science and society))*
5. 1 or 2 or 3 or 4
6. 5 and (science or biology or chemistry or physics or earth science)
7. 6 and (course or courses or program$ or project$ or material$ or curricul$ or intervention$)
8. limit 7 to (English language and (elementary secondary education or elementary education or intermediate grades or secondary education or middle schools or junior high schools or high schools or high school equivalency programs or postsecondary education or two year colleges) and (books or ‘collected works (general and serials)’ or conference proceedings or dissertations or ‘evaluative or feasibility reports’ or general reports or journal articles or project descriptions or ‘research or technical reports’ or ‘speeches or conference papers’) and yr=1980-2002)

1b British Education Index (BEI)

BEI was searched on 11 February 2003, using the BIDS Ovid interface. The database was searched for the period 1980–2002 and 176 records were retrieved.

1. (context-based or context-led or application-based or application-led)
2. (contextualis$ or contextualiz$ or technological approach)
3. (Salters or SATIS or PLON or ChemCom or Science in context)
4. (science-technology-society or (science and society))
5. (out-of-school experience or everyday or authentic)
6. 1 or 2 or 3 or 4 or 5 or 6 and (science or biology or chemistry or physics or earth science)
7. limit 7 to English language

1c PsycINFO

PsycINFO was searched on 12 February 2003, using the WINSPIRS interface. The database was searched for the period 1980 – 2002 and 33 records were retrieved.

1. context-based or context-led or application-based or application-led
2. contextualis* or contextualiz* or technological approach
3. Salters or SATIS or PLON or ChemCom
4. science-technology-society or science-technology-society or sciencetechnologysociety or science-through-technology or socioscientific or socio-scientific
5. #1 or #2 or #3 or #4
6. #5 and (education* or school* or college or student* or pupil* or learner*) and (science or biology or chemistry or physics or earth science)
7. #5 and (education* or school* or college or student* or pupil* or learner*) and (science or biology or chemistry or physics or earth science)
8. #7 and (course or courses or program* or project* or curricul* or material* or intervention*)
9. #8 and (LA:PY = English) and ((PT:PY = case-study) or (PT:PY = clinical-trial) or (PT:PY = collected-works) or (PT:PY = conference-proceedings-symposia) or (PT:PY = empirical-study) or (PT:PY = experimental-replication) or (PT:PY = followup-study) or (PT:PY = interview) or (PT:PY = journal-abstract) or (PT:PY = literature-review-research-review) or (PT:PY = longitudinal-study) or (PT:PY = meta-analysis) or (PT:PY = program-evaluation) or (PT:PY = prospective-study) or (PT:PY = retrospective-study) or (PT:PY = treatment-outcome-study)) and (PY:PY = 1980–2002)

1d. Social Science Citation Index

The Social Science Citation Index was searched on 12 February 2003, using the Web of Science interface. The database was searched for the period 1992–2003 and 210 records were retrieved. It was not possible to download search strings with the records, but these largely replicated those used for ERIC and BEI.

2. Handsearches

The journal *Studies in Science Education* was handsearched as it contains major review articles and was seen as a potential key source of studies.

Appendix 2.3: EPPI-Centre keyword sheet, including review-specific keywords

V0.9.7 *Bibliographic details and/or unique identifier*

A1. Identification of report
Citation
Contact
Handsearch
Unknown
Electronic database
(Please specify.) ....

A2. Status
Published
In press
Unpublished

A3. Linked reports
Is this report linked to one or more other reports in such a way that they also report the same study?

Not linked
Linked (Please provide bibliographical details and/or unique identifier.)
....
....
....
....

A4. Language (Please specify.)
....

A5. In which country/countries was the study carried out? (Please specify)
....
....
....

A6. What is/are the topic focus/foci of the study?
Assessment
Classroom management
Curriculum*
Equal opportunities
Methodology
Organisation and management
Policy
Teacher careers
Teaching and learning
Other (Please specify.)........................

A7. Curriculum
Art
Business studies
Citizenship
Cross-curricular
Design and technology
Environment
General
Geography
Hidden
History
ICT
Literacy – first language
Literacy further languages
Literature
Maths
Music
PSE
Physical education
Religious education
Science
Vocational
Other (Please specify.)........................

A8. Programme name (Please specify.)

....

A9. What is/are the population focus/foci of the study?
Learners
Senior management
Teaching staff
Non-teaching staff
Other education practitioners
Government
Local education authority officers
Parents
Governors
Other (Please specify.)...........................

A10. Age of learners (years)
0–4
5–10
11–16
17–20
21 and over

A11. Sex of learners
Female only
Male only
Mixed sex

A12. What is/are the educational setting(s) of the study?
Community centre
Correctional institution
Government department
Higher education institution
Home
Independent school
Local education authority
Nursery school
Post-compulsory education institution
Primary school
Pupil referral unit
Residential school
Secondary school
Special needs school
Workplace
Other educational setting (Please specify.)...

A13. Which type(s) of study does this report describe?
A. Description
B. Exploration of relationships
C. Evaluation
   a. naturally-occurring
   b. researcher-manipulated
D. Development of methodology
E. Review
   a. Systematic review
   b. Other review

Review-specific keywords

For each item tick any number of keywords **1. What discipline?** a. (integrated) Science b. Biology c. Chemistry d. Physics e. Earth Science f. Environmental Science **2. What types of learners are involved?** a. mixed ability b. lower ability/slow learners c. middle ability d. upper ability/gifted e. disaffected f. non-science	**3. What is the nature of the intervention?** a. full context-based course b. context-based unit/module c. full STS course d. STS enrichment topics **4. What aspect of science learning does the intervention focus on improving?** a. Attitudes *(e.g. attitude to science, attitude to school science, motivation to learn, interest in science activities, social/group collaboration, career intentions)* b. Understanding of science *(e.g. science concepts, the nature of science, scientific methods)* c. Skills *(e.g. investigative skills, manipulative skills, communication skills, problem-solving skills, decision-making skills)*	**5. How are outcomes measured?** a. test results b. external examination results c. written reports/open questionnaires d. concept webs e. (dis)agreement scores (including VOSTS) f. self-reports *(e.g. diaries, interviews)* g. recorded group discussions h. presentations i. observed behaviour **6. Outcomes report on effectiveness in terms of:** a. understanding of science b. attitudes to science c. skills **for learners of different:** d. ability (lower/middle/higher) e. attitude (including disaffected) f. science specialisation (including non-specific) g. gender h. educational level

Appendix 2.4.1: Indicators for weight of evidence – gender and attitude (GA)

Review question (GA): What is the evidence from evaluative studies of the effect of context-based or STS courses on the attitude to science of boys and girls in the 11 to 16 age range?

Weight of evidence B Appropriateness of research design and analysis for addressing the question *of this specific systematic review*			**Weight of evidence C** Relevance of particular focus of the study (incl. conceptual focus, context, sample and measures) for addressing the question *of this specific systematic review*			**Weight of evidence D** Taking into account A, B and C, what is the overall weight of evidence this study provides to answer *this review question*?
For the RQs relevant to the review… The study is an evaluation of gender effect. If not, final weight for B: LOW If so, weighting according to aspects below			***For the RQs relevant to the review…***			If equal weighting of A, B and C, each weighted across the range as low (1), medium low (2), medium (3), medium high (4) and high (5) Sum total and classification for D: 3–4: low 5–7: medium low 8–10: medium 11–13: medium high 14–15: high
high (3)	**medium (2)**	**low (1)**	**high (3)**	**medium (2)**	**low (1)**	
sample size			**nature of sample**			
large sample with appropriate sampling method	large sample, no sampling method	small sample (up to three classes)	balanced gender ratios during intervention	unbalanced gender ratios during intervention	single sex groups during intervention	
comparison/control			**focus of intervention**			
comparison for gender in design (control)	comparison for gender in findings only	no comparison/control	gender is explicit independent variable	gender is major discrete element of intervention	gender wrapped up in intervention	
benchmark data			**measures**			
pre-post data on attitude to science	longitudinal dev of attitude to science	only post data for attitude to science	highly appropriate for testing directly attitude to science	mildly appropriate for testing directly attitude to science	appropriate for testing attitude to science indirectly	
data-collection			**breadth**			
solid checks on reliability/validity for data-collection	some checks on reliability/validity for data-collection	little/no checks on reliability/validity for data-collection	reports broad range of attitudes to science	reports narrow range of attitudes to science	only reports attitudes to science indirectly	
data analysis			**situation**			
solid checks on reliability/validity for data analysis	some checks on reliability/validity for data-collection	little/no checks on reliability/validity for data-collection	highly representative of learners in classrooms	less representative of learners in classrooms	not in classrooms	

For both B and C: totals 5–6=low; 7–8=medium low; 9–11=medium; 12–13=medium high; 14–15=high

Appendix 2.4.2: Indicators for weight of evidence – gender and understanding (GU)

Review question (GU): What is the evidence from evaluative studies of the effect of context-based or STS courses on the understanding of science ideas of boys and girls in the 11 to 16 age range?

Weight of evidence B Appropriateness of research design and analysis for addressing the question *of this specific systematic review*			**Weight of evidence C** Relevance of particular focus of the study (incl. conceptual focus, context, sample and measures) for addressing the question *of this specific systematic review*			**Weight of evidence D** Taking into account A, B and C, what is the overall weight of evidence this study provides to answer *this review question*?
For the RQs relevant to the review… The study is an evaluation of gender effect. If not, final weight for B: LOW If so, weighting according to aspects below			***For the RQs relevant to the review…***			If equal weighting of A, B and C, each weighted across the range as low (1), medium low (2), medium (3), medium high (4) and high (5) Sum total and classification for D: 3–4: low 5–7: medium low 8–10: medium 11–13: medium high 14–15: high
high (3)	**medium (2)**	**low (1)**	**high (3)**	**medium (2)**	**low (1)**	
sample size			**nature of sample**			
large sample with appropriate sampling method	large sample, no sampling method	small sample (up to three classes)	balanced gender ratios during intervention	unbalanced gender ratios during intervention	single sex groups during intervention	
comparison/control			**focus of intervention**			
comparison for gender in design (control)	comparison for gender in findings only	no comparison/control	gender is explicit independent variable	gender is major discrete element of intervention	gender wrapped up in intervention	
benchmark data			**measures**			
pre-post data on science understanding	longitudinal dev of science understanding	only post data for science understanding	highly appropriate for testing directly science understanding	mildly appropriate for testing directly science understanding	appropriate for testing science understanding indirectly	
data-collection			**breadth**			
solid checks on reliability/validity for data-collection	some checks on reliability/validity for data-collection	little/no checks on reliability/validity for data-collection	reports broad range of science understanding	reports narrow range of science understanding	only reports science understanding indirectly	
data analysis			**situation**			
solid checks on reliability/validity for data analysis	some checks on reliability/validity for data-collection	little/no checks on reliability/validity for data-collection	highly representative of learners in classrooms	less representative of learners in classrooms	not in classrooms	

For both B and C: totals 5–6=low; 7–8=medium low; 9–11=medium; 12–13=medium high; 14–15=high.

Appendix 2.4.3: Indicators for weight of evidence – lower ability and attitude (LA)

Review question (LA): What is the evidence from evaluative studies of the effect of context-based or STS courses on the attitude to science of lower-ability pupils in the 11 to 16 age range?

Weight of evidence B Appropriateness of research design and analysis for addressing the question *of this specific systematic review*			**Weight of evidence C** Relevance of particular focus of the study (incl. conceptual focus, context, sample and measures) for addressing the question *of this specific systematic review*			**Weight of evidence D** Taking into account A, B and C, what is the overall weight of evidence this study provides to answer *this review question*?
For the RQs relevant to the review… The study is an evaluation of the learning effect on low-ability learners. If not, final weight for B: LOW If so, weighting according to aspects below			***For the RQs relevant to the review…***			If equal weighting of A, B and C, each weighted across the range as low (1), medium low (2), medium (3), medium high (4) and high (5) Sum total and classification for D: 3–4: low 5–7: medium low 8–10: medium 11–13: medium high 14–15: high
high (3)	**medium (2)**	**low (1)**	**high (3)**	**medium (2)**	**low (1)**	
sample size			**nature of sample**			
large sample with appropriate sampling method	large sample, no sampling method	small sample (up to two classes)	highly representative of low-ability learners	less representative of low-ability learners	not representative of low-ability learners	
comparison/control			**focus of intervention**			
comparison/control for low ability in design	comparison for low ability in findings only	no comparison/control	low ability is explicit independent variable	low ability is discrete element of intervention	low ability is wrapped up in intervention	
benchmark data			**measures**			
pre-post data on attitude to science	longitudinal dev of attitude to science	only post data for attitude to science	highly appropriate for testing directly attitude to science	mildly appropriate for testing directly attitude to science	appropriate for testing attitude to science indirectly	
data-collection			**breadth**			
solid checks on reliability/validity for data-collection	some checks on reliability/validity for data-collection	little/no checks on reliability/validity for data-collection	reports broad range of attitudes to science	reports narrow range of attitudes to science	only reports attitudes to science indirectly	
data analysis			**situation**			
solid checks on reliability/validity for data analysis	some checks on reliability/validity for data-collection	little/no checks on reliability/validity for data-collection	highly representative of learners in classrooms	less representative of learners in classrooms	not in classrooms	

For both B and C: totals 5–6=low; 7–8=medium low; 9–11=medium; 12–13=medium high; 14–15=high.

Appendix 2.4.4: Indicators for weight of evidence – lower ability and understanding (LU)

Review question (LU): **What is the evidence from evaluative studies of the effect of context-based or STS courses on the understanding of science ideas of lower-ability pupils in the 11 to 16 age range?**

Weight of evidence B Appropriateness of research design and analysis for addressing the question *of this specific systematic review*			**Weight of evidence C** Relevance of particular focus of the study (incl. conceptual focus, context, sample and measures) for addressing the question *of this specific systematic review*			**Weight of evidence D** Taking into account A, B and C, what is the overall weight of evidence this study provides to answer *this review question*?
For the RQs relevant to the review… The study is an evaluation of the learning effect on low-ability learners. If not, final weight for B: LOW If so, weighting according to aspects below			***For the RQs relevant to the review…***			If equal weighting of A, B and C, each weighted across the range as low (1), medium low (2), medium (3), medium high (4) and high (5)
high (3)	**medium (2)**	**low (1)**	**high (3)**	**medium (2)**	**low (1)**	Sum total and classification for D: 3–4: low 5–7: medium low 8–10: medium 11–13: medium high 14–15: high
sample size			**nature of sample**			
large sample with appropriate sampling method	large sample, no sampling method	small sample (up to two classes)	highly representative of low-ability learners	less representative of low-ability learners	not representative of low-ability learners	
comparison/control			**focus of intervention**			
comparison/control for low ability in design	comparison for low ability in findings only	no comparison/control	low ability is explicit independent variable	low ability is discrete element of intervention	low ability is wrapped up in intervention	
benchmark data			**measures**			
pre-post data on science understanding	longitudinal dev of science understanding	only post data for science understanding	highly appropriate for testing directly science understanding	mildly appropriate for testing directly science understanding	appropriate for testing science understanding indirectly	
data-collection			**breadth**			
solid checks on reliability/validity for data-collection	some checks on reliability/validity for data-collection	little/no checks on reliability/validity for data-collection	reports broad range of science understanding	reports narrow range of science understanding	only reports science understanding indirectly	
data analysis			**situation**			
solid checks on reliability/validity for data analysis	some checks on reliability/validity for data-collection	little/no checks on reliability/validity for data-collection	highly representative of learners in classrooms	less representative of learners in classrooms	not in classrooms	

For both B and C: totals 5–6=low; 7–8=medium low; 9–11=medium; 12–13=medium high; 14–15=high.

Appendix 4.1: Summary tables of studies included in the in-depth review

Ebenezer JV, Zoller U (1993) Grade 10 students' perceptions of and attitude toward science teaching and school science

Country of study	Canada: British Columbia
Details of researchers	Researchers from the Universities of Manitoba and Haifa, Israel
Name of programme	British Columbia Junior Secondary Science Program (JSSP)
Age of learners	11–16, Grade 10
Type of study	Naturally-occurring evaluation

Aims of study

- To assess the perceptions of Grade 10 students concerning the JSSP science instruction they received
- To assess their attitudes to school science
- To identify factors influencing their perceptions of and their attitudes towards the science they learn in school

Summary of study design, including details of sample

- Large-scale survey of student perceptions
- Target population: all Grade 10 students in BC in 1989 cohorts
- Selected sample questionnaires: all Grade 10 students in one district (selected for convenient access and representativeness for BC schools). Actual sample: 1,094 (540 girls, 520 boys, 34 NI by gender; 69.9% of sample) in 1989. No pre-intervention data, but comparison (for attitude to school science) with similar pre-intervention 1986 cohort of 1,316 students (no gender breakdown given).
- Selected sample for interviews: 72 students stratified for positive, neutral and negative attitudes to school science emerging from the questionnaire responses, and controlled for gender and for schools (six). Actual sample: only 32 students (13 boys, 19 girls) were able to participate; two schools are under-represented.

Methods used to collect data

- Likert-scale questionnaires administered to each cohort
- Audiotaped, semi-structured 10- to 30-minute interviews

Data-collection instruments, including details of checks on reliability and validity

- Questionnaires with 29 items, of which nine relate to perceptions of frequencies of classroom activities taking place in a standard week (rated as always, often, sometimes, never) and seven to statements of attitudes to school science (five-point-scale SA to SD). Based on BC Science Assessment Instrument used in 1986 study. Text of each item is provided. Factor analysis used for validating scales (Hoyt reliability estimate 0.89). Remainder of 29 items not specified.
- Interview schedule provided: contains very general prompts (mostly positive). No mention of validity or reliability checks for interviews. However, the data from the interviews were used 'to complement the data that were obtained from the previously described components of the assessment instrument and facilitated a better interpretation of the results'. 'The depth interview method (Jones, 1985) was employed to validate the positions (positive, neutral, or negative) taken by students in terms of their attitudes toward science teaching. Additionally, detailed information was elicited on students' perceptions of the science teaching done by their teachers.'

Methods used to analyse data, including details of checks on reliability and validity

- For perceptions of activities: Frequencies of Likert scale choices (as sample percentage) calculated and presented in a table. No across-students weighting. Standard problems with validity of aggregating Likert scale scores. For half the items, some gender distribution is provided, but no significance testing. Much of the gender breakdown is impossible, or implausible.

- For attitudes to school science: Responses were aggregated to give three categories as positive (strongly agree and agree), neutral (undecided) and negative (strongly disagree and disagree) about science statements. Frequencies (as sample percentage) were calculated and presented in a table. No across-students weighting. For half the items some gender distribution is provided. Some informal reliability checks by inspection of response distribution on similar items. Weighted aggregated scores per student for attitude across items, with resulting percentage of sample's attitude as positive, neutral or negative (presented in table, including comparison for 1996 cohort). Mean scores for 1989 and 1986 are compared. Weighting method unclear, with z-value mentioned in text for significance of gender differences.
- Interview transcripts used for validating questionnaire responses (on attitude) and elaborate perceptions of classroom activities. Five quotations provided. Analysis method for interviews not specified, but it looks to be simple grounded theory method. No triangulation for individual students attempted. No mention of independent coding of responses to increase reliability.

Summary of results

Note: Many results are reported as facts, whereas the data only allow for students' perceptions. For several of the items, no gender information was provided.

Perceptions of predominant classroom activities (Table 2):

- A large majority of students (87.5%, no gender difference) report that they frequently copy teacher's notes with, in addition, just less than half (more of the boys) claiming that the teacher frequently hands out notes.
- Extensive note-taking, reading text with answering comprehension questions, and memorisation without understanding makes students dislike science learning (from interviews).
- A majority (52.5%) reports working frequently in small groups; authors report this as small group practical work.
- A majority of students report hardly being involved in individual practical work (73.7%), and never using the computer (89.4%).
- Many students (66.2%, more girls) report that their ideas were never used when planning science lessons.
- Student-directed experimentation and explorations and the study of science related to everyday life are favourite activities (from interviews).

Attitude to science:

- More than half (58.0%) the students report they like studying science and 72.9% said that studying school science is important. However, only 38.3% wish to study more science. Gender distributions given are impossible or improbable.
- In aggregate, just over 40% of students feel positive, around 45% are neutral, and 13% feel negative towards school science. No gender differences here, and no differences with the 1986 cohort.

Conclusions

'Students place a great deal of emphasis on science teaching practice and its relationship to the enjoyment and comprehension of science. ... Students wished to be taught from their own perspectives; that is, by relating school science to everyday science in the science classroom. …Most noteworthy is the fact that the type of teaching students prefer is consistent with the curricular changes promoted by constructivist and STS educators' (p 184).

'In this study teaching style played a central role in determining student' attitudes toward science' (no evidence provided).

Note: Authors do not comment in their conclusions on any of the gender differences emerging.

Weight of evidence A (trustworthiness in relation to study questions)

Medium: Large sample, supposedly representative. No pre-post design but (unconvincing) comparison with previous cohort (for attitude to school science only). Balanced for gender, but no information on nature/reason for 30% non-responses. Co-ordinated quantitative and qualitative design. Use of Likert scales with associated validity problems for aggregation. No piloting. Only frequency statistics, no inter-item reliability checks, or calculation for significant differences for gender. No analysis method provided for interview data, no reliability/validity checks.

Weight of evidence B (GA) (appropriateness of research design and analysis)

Overall low: The study is not an evaluation of gender effect.

Weight of evidence C (GA) (relevance of focus of study to review)

Medium: Balanced gender ratios for questionnaires and reasonable balance for interviews; gender not explicit focus of intervention; some measures (Likert scales) tested attitude, others reported activities only; limited range in items (four issues addressed only) on attitude to school science; situation is representative for learners in schools.

Weight of evidence D (GA)

Medium low

Greenlee S, Lambert L (1996) The Science of Living Spaces: Women in the Environment of the 21st Century. An Evaluation and Guide Book

Country of study	USA
Details of researchers	Researchers at the University of Christopher Newport, Virginia
Name of programme	Science of Living Spaces: Women in the Environment of the 21st Century
Age of learners	11–13
Type of study	Naturally-occurring evaluation

Aims of study

- The programme being evaluated was designed to encourage girls to consider careers in science, engineering or mathematics (SEM). The study was a year long but the core and most intensive part was a three-week, summer, residential programme.
- The aim of the evaluation was to assess the project’s success, prepare a guide book on the implementation of the programme, disseminate information about the programme and refine the programme.

Summary of study design, including details of sample

- The evaluation was multi-faceted and used both quantitative and qualitative assessment techniques, and a pre-post test design. Standardised measures were used to assess general aspects and content, and project-specific measures used to assess the particular content of the project activities.
- Sample was 24 girls aged 11–13 (7th and 8th grade) with a mean age of 11.9. Demographic data are supplied in Table 1. The sample was selected from interested girls applying to join the programme. Average to above average girls ability were chosen, all of whom wanted to go to college. A wide range of socio-economic backgrounds was represented.

Methods used to collect data

Quantitative and qualitative. Standardised generic measures and project-specific measures

Data-collection instruments, including details of checks on reliability and validity

Three standardised measures were used.

- Self-concept was measured using the Tennessee Self-Concept Scale (TSCS).
- Academic motivation was measured using the Children’s Academic Intrinsic Motivation Inventory (CAIMI).
- Career interest patterns were assessed using the Self-Directed Search Career Explorer (SDSCE).

Five project-specific measures were used:

- One used Demographic Descriptive Profile Assessment, which measures interest in careers, attitudes, interests and awareness.

- The Knowledge Skills Building Assessment is a self-administered, multiple-choice questionnaire used to assess knowledge of and exposure to project-specific content. These were developed by project leaders.
- The Parent/Teacher Assessment (P/TA) is an assessment instrument developed to obtain parents' and teachers' assessment of the goals and abilities. Parents rate and assess their daughters' behaviour, feelings and attitudes towards science, and also rate their interests in science mathematics and engineering.
- Teachers (using the P/TA) were asked to assess participants' maturity, study skills, problem-solving skills, ability to think independently, and excitement about learning.
- The General Assessment attempted to obtain feedback on specific aspects of activities and projects in the programme and involved giving participants a simple evaluation sheet at the end of each week of the programme and during the residential week. No mention of reliability or validity checks for these.

Details of the tests are given in Appendices of the paper.

Methods used to analyse data, including details of checks on reliability and validity

- Used standard tests – mean, standard deviation and t-tests to analyse the pre-post test data. Likert-scaled responses were presented as means and percentages. No statistical analysis of these were given.
- Where response levels were low, data were not reported.
- It could be argued that multiple data sources enhanced reliability, and using previously validated standardised tests enhanced this validity.

Summary of results

- The training session for parents rated 4.48 on a five-point Likert scale. 94.5% of parents reported that their training raised their awareness of the factors that impact on their daughters' selection of careers in SEM (Table 2).
- Training session for staff received an overall rating of 4.63 on a five-point Likert scale (Table 3).
- CAIMI test showed a slight increase on girls interest in science and mathematics but not significant, probably due to selected sample already being interested in these subjects (Table 4). 96% of girls said that they were extremely interested in taking more science and mathematics courses, and 96% indicated that they did like science more as a result of the Science Living Spaces programme.
- Self-concept test showed a slight increase in level of self-esteem but not statistically significant (Table 5). Parents rated their daughters' self-esteem significantly higher after participating in the programme.
- Career interests pre-post scores indicated an increase on the Investigative (science) scale but not the Conventional (mathematics) scale. Differences were not significantly different (Table 6).
- Enjoyment of programme average to above average ratings were given for each activity on a weekly basis (Table 7).
- Knowledge/skills were *not* reported on.
- Anecdotal qualitative responses are reported from girls, staff and parents, and all were 'generally very positive'.

Conclusions

Authors report that the programme was extremely successful in meeting its goals and objectives: 'The participants and their parents reported a greater interest in science related careers' and girls reported that meeting 'women scientists played a significant role in helping them to overcome some of the stereotypic views and encouraged them to look more positively at the contributions of women in sciences'. However, in the four cases where statistical information was given, it did not support the view that significant changes had been made.

The other main conclusion was that the programme was worth re-running with some adjustments.

Weight of evidence A (trustworthiness in relation to study questions)

Medium low: A positive feature is that a wide range of instruments was used to gather data. However, few details are provided of reliability and validity checks for the project specific measures developed. Used statistical tests for four measures, found no statistical difference but still claimed that the programme was a success based on anecdotal, qualitative responses. Not all aspects measured were reported (knowledge gained).

Weight of evidence B (GL) (appropriateness of research design and analysis for review)

Medium low: Small sample size; no control for gender; used pre-post tests for some measures; some checks on reliability and validity by using published tests for some measures but not for all; used standard statistical tests for quantitative data and used triangulation of views from girls, teachers and parents for qualitative evaluation of the programme generally.

Weight of evidence C (GL) (relevance of focus of study to review)

Medium: Sample not very representative as chosen for interest in science and science career, and average or higher than average ability; gender major focus of study; measures were good for testing attitude to science; broad range of measures used; situation not at all representative (major part was a three-week residential course at university).

Weight of evidence D (GL) (overall weight of evidence)

Medium low

Huppert *et al*. (1992) Human health and science: a model for an STS high school biology course

Country of study	Israel
Details of researchers	Researchers from Haifa University and Haifa Technion, and one school teacher
Name of programme	Human Health and Science (HH&S), context-based intervention at Grades 11 and 12, aimed at (i) teaching biological concepts and processes in a relevant context, (ii) highlighting the connection between SandT in health care; (iii) creating awareness of the social and moral consequences of science progress; and (iv) enhancing pupils' motivation to learn. Course consists of five independent modules often taught simultaneously. Biology concepts are supplemented with technology applications and moral issues, using a diversity of teaching methods.
Age of learners	11–16, but age not provided explicitly. Assumed post-Grade 10, as Grade 10 mathematics scores were used to assess ability levels.
Type of study	Naturally-occurring evaluation

Aims of study

- What are the pupils' achievements due to the intervention? (RQ1)
- What are the pupils' attitudes towards the intervention? (RQ2)

Summary of study design, including details of sample

- Small-scale evaluation by pilot class
- Post-intervention data only
- Sample consists of one class only (20 girls, 15 boys, 15 under-achievers) on an elective course. No selection method reported.

Methods used to collect data

- Self-constructed, end-of-course achievement test
- Likert-scale attitude questionnaire

Data-collection instruments, including details of checks on reliability and validity

- The achievement test contained 15 multiple-choice questions, three open questions, comprehension task based on journal paper and film. Content validity checked by five teachers. No test of equivalence with tests of traditional courses.
- The attitude questionnaire had 10 Likert-scale items, five positive, five negative, clustered in four course objective areas. Content validation by five teachers. Reliability checked by Alpha Cronbach test (0.79).

Methods used to analyse data, including details of checks on reliability and validity

- Achievement test: Scoring of correctness of responses per student. Calculating the percentage mean. No indication of independent marking of open and comprehension responses. No statistical comparisons.
- Attitude test: Rated responses aggregated and sample's mean and standard deviation (SD) calculated per item (provided in Table 1), and the same for the four objective areas (Table 2).

Summary of results

- Achievement: Mean achievement of 79.5% in (HH&S), versus 67.0% in former (traditional) biology course. Claim (p 398) that 'it can be seen that students with low learning motivation and low academic achievement were more successful when instructed in an STS approach' (but no evidence is provided).
- Attitude: (from Table 2) High appreciation of all four course objectives (maximum 4.00), but highest for awareness of social/moral consequences of science (mean 3.30, SD 0.73) and learning biology in context (mean 3.15, SD 0.73). Lowest appreciation of motivation to learn (mean 2.76, SD 0.69).

Conclusions

Results repeat findings.

Weight of evidence A (trustworthiness in relation to study questions)

Low: Very small sample size, no comparable baseline data. Achievement and attitude tests checked for content validity. Standard query of validity of aggregating Likert-scale responses. No independent marking of open responses. Claims concerning low achievers not supported by data.

Weight of evidence B (LA) (appropriateness of research design and analysis)

Low: Very small sample size (35), no sampling method; no benchmark data; some check on validity but none for reliability of data-collection tool; standard quantitative treatment (mean/SD) but standard problems with Likert-scale responses.

Weight of evidence C (LA) (relevance of focus of study to review)

Medium: No detail on type of low-ability learners, only percentage within group; low ability is one of the variables, but no data are given to compare change of view; reasonable measures for ability; breadth of attitudes covering the objectives of course, but provides limited scope for attitudes to science; highly representative of elective classrooms.

Weight of evidence D (LA) (overall weight of evidence)

Medium low

Weight of evidence B (LU) (appropriateness of research design and analysis)

Low: Very small sample size (35); no sampling method; benchmark data based on general previous performance in biology; some check on validity, but none for reliability of data-collection tool; no detail on analysis method.

Weight of evidence C (LU) (relevance of focus of study to review)

Medium: No detail on type of low-ability learners, only percentage within group; low ability is one of the variables, but no data are given to compare change of view; reasonable measures for ability; breadth of attitudes covering the objectives of course, but provides limited scope for attitudes to science; highly representative of elective classrooms.

Weight of evidence D (LU) (overall weight of evidence)

Medium low

Nentwig *et al.* (2002) Chemie im Kontext: from situated learning in relevant contexts to a systematic development of basic chemical concepts

Country of study	Germany
Details of researchers	Not explicit but presume developers of the programme at various universities. Data for evaluation studies are being collected by doctoral students in some cases; in others, it is not clear.
Name of programme	Chemie im Kontext (ChiK)
Age of learners	Upper secondary – ages not given
Type of study	Naturally-occurring evaluation

Aims of study

- The broad aims of the studies were to explore the effects of ChiK on pupils' motivation for learning chemistry, on their understanding of chemical concepts, and on the application of chemical knowledge to real-life contexts. The evaluation was carried out through two research questions.
- Does the course influence the learners' perceptions of the instructional quality? (RQ1)
- What are the cognitive learning outcomes of the course? (RQ2)

This report gives data on the first RQ but the study for RQ2 is still ongoing and is described in more general terms.

Summary of study design, including details of sample

- Pre-post study of learners' perceptions of instruction (RQ1).
- Case studies to investigate outcomes in detail regarding course content; basic chemical concepts, scientific concepts in relation to daily life, and transferability of concepts and meta-cognition (RQ2).
- 110 pupils from eight classes in five schools in Upper Saxony. No details of schools, gender, ability or socio-economic background.

Methods used to collect data

- Quantitative (RQ1)
- Case studies – ongoing; described but no findings reported (RQ2).

Data-collection instruments, including details of checks on reliability and validity

- A 54-item questionnaire of learners' perceptions of a number of criteria of quality of T and L using five-step scale (?Likert scale). Pupils were given the tests before and after each Chemie im Kontext unit (15–20 lessons) (RQ1).
- Very little detail given for RQ2. Three-part test instrument: (a) problem-solving, based on course content, (b) understanding of chemical concepts and (c) problem-solving in new context. Other information collected on student changing conceptual understanding and insights into learning but limited detail given in this report.

Methods used to analyse data, including details of checks on reliability and validity

- Factor analysis was used to aggregate the 54 questions into five variables: interest (six items), relevance (seven items), own competence (five items); instructional quality (six items) and learning atmosphere (seven items). α ratings of between 0.74 and 0.86 quoted for validity. t-test was used for pre-post differences using paired samples. Looked for gender and ability differences. Ability (successful/non-successful) was assigned in terms of previous grades in chemistry) (RQ1).
- N/A as data for RQ2 not yet complete.

Summary of results

- Significant positive effects for perception of relevance of chemical contents and for quality of instruction $p<0.001$.

- Some interaction effects but quantitative data not given: pupils with lower chemistry grades in previous school year showed higher increase in all variable of T and L than pupils with better grades.
- No significant differences were found with gender.

Conclusions

Chemie im Kontext provides teachers with conceptual framework in creating different chemistry classes which consider goals of science classes, motivation and cognitive processes. The authors conclude that teachers' worries about lower-ability pupils coping with the demands of the course appeared to be unfounded on the basis of pupils' own reported perceptions.

Weight of evidence A (trustworthiness in relation to study questions)

Low: This is a pilot study and few basic data are given. Much of the paper is taken up with describing the course and its development under the influence of Salters Advanced Chemistry and ChemCom and plans for future development.

Weight of evidence B (GA) (appropriateness of research design and analysis)

Low: This was not an evaluation of gender.

Weight of evidence C (GA) (relevance of focus of study to review)

Medium: Balance of gender not given but included both; gender was one of the explicit variables; measures for RQ1 (perception of the quality of instruction appropriate) but little detail given for RQ2 measures; breadth of attitude questions good in RQ1 but no details given for RQ2; situation would seem to be representative of learners in classrooms but not much detail given.

Weight of evidence D (GA) (overall weight of evidence)

Medium low

Weight of evidence B (LA) (appropriateness of research design and analysis)

Low: This was not an evaluation of responses of pupils of different abilities.

Weight of evidence C (LA) (relevance of focus of study to review)

Medium: No detail given of balance of low-ability pupils; low ability is one of the explicit variables; measures for RQ1 (perception of the quality of instruction appropriate) but little detail given for RQ2 measures; breadth of attitude questions good in RQ1 but no details given for RQ2; situation would seem to be representative of learners in classrooms but not much detail given.

Weight of evidence D (LA) (overall weight of evidence)

Medium low

Ramsden, JM (1992) If it's enjoyable, is it science?

Country of study	UK: England
Details of researchers	Researcher from University of York
Name of programme	Salters' Science
Age of learners	13–14
Type of study	Naturally-occurring evaluation

Aims of study

- What are the reactions of Y9 pupils to the Salters' course? (Explicit)
- How do these reactions differ for gender? (Additional implicit)

Summary of study design, including details of sample

- Post-intervention survey
- Six schools were selected from approximately 100 trial schools, if at least eight unaltered context-based units were used in Y9, with completely mixed-ability and mixed-sex groups.

- Sample was 124 pupils (59 girls, 65 boys) from the six schools.

Methods used to collect data

- Pupils from one class per school answered a questionnaire about their views of the course at the end of the academic year, using short descriptions of the units if needed expanded by the teacher.

Data-collection instruments, including details of checks on reliability and validity

- Six Likert-scale (five points, from strongly agree to strongly disagree) type statements asked for views on the course.
- One open-ended question asking for elaboration of the choices.
- No mention of reliability checks (although author notes that all items are positive).
- The validity was improved by providing short descriptions of each unit with an expansion by the teacher if required. The open-ended responses provided an opportunity for validation, but no analysis reported.

Methods used to analyse data, including details of checks on reliability and validity

- Mean values (no SD) for responses per gender were calculated (Table 2) for each of the six items, only used for comparison of overall ratings for different units, of ratings per gender for each unit, or of overall ratings across all units for each item.
- Standard problem of validity with aggregation of these Likert scores.
- Significant differences calculated (critical ratio analysis of mean values and χ^2 method for frequencies). Responses to open-ended questions simply grouped and frequencies determined.
- No mention of blind coding of these responses.
- Potential bias, since teachers in trial schools will be more positively inclined to the chosen approach than the average teacher.

Summary of results

- In terms of appeal of the context-based approach, there were few differences between girls and boys suggesting that the activities produced a similar response in members of both sexes.
- Where boys enjoyed a unit significantly more than girls (two out of ten units), this was either because of the nature of the practical work in the unit or on account of the resulting increase of interest in science.
- Where girls enjoyed a unit significantly more than boys (two out of ten units), this was either because of the non-practical activities in the units or on account of the perceived usefulness in later life.
- All units attracted favourable comments from both boys and girls.
- Pupils' enjoyment of a unit did not correlate with an increased interest in science. The latter was rated significantly lower.
- Boys enjoyed practical work significantly more than non-practical work (mainly seen as writing). No difference in enjoyment emerged for girls.

Conclusions

Interest in science lessons may increase by using everyday interests as a starting point and using a wide range of learning activities. A mismatch exists in pupils' minds between context-based activities and learning science.

Weight of evidence A (trustworthiness in relation to study questions)

Medium: Research questions not very specific. Reasonable sample size with carefully controlled contexts (mixed ability and mixed gender). Bias possible since these are trial schools. Instrument items not validated or piloted, but open-ended responses validated the Likert scores. Standard problem with aggregation of Likert responses, but limited to use for comparisons, no significance attached to absolute scores. Standard statistics used appropriately for significance testing, but not reported numerically. No validity checks for judgements of open-ended responses.

Weight of evidence B (GA) (appropriateness of research design and analysis)

Medium: Reasonable sample size with sampling method (but trial schools introduce bias); gender balanced in sample; no benchmark data; no pilot, some reliability/validity check for data-collection through summary of all units; standard statistics for data analysis of Likert responses, but no validity check for open-ended responses.

Weight of evidence C (GA) (relevance of focus of study to review)

High: Balanced gender ratio (mixed sex classes only); gender explicit variable in analysis (not in research question); Likert-scale statements plus elaboration appropriate for testing attitude; covers broad range of attitudes to school science (activities, materials, learning outcomes, interest and usefulness, but not career choice); situation (mixed ability, mixed gender) highly representative.

Weight of evidence D (GA) (overall weight of evidence)

Medium high

Ramsden JM (1997) How does a context-based approach influence understanding of key chemical ideas at 16+?

Country of study	UK: England
Details of researchers	Researcher from University of York
Name of programme	Science: The Salters' Approach
Age of learners	15–16
Type of study	Researcher-manipulated evaluation

Aims of study

- To explore the effect of a context-based approach on pupils' understanding of some key chemical ideas at 16+
- To explore pupils' views on enjoyment of lessons and career intentions

Summary of study design, including details of sample

- Experiment – intervention group and control group
- Sample of 168 pupils in eight schools, four using context-based course and four using conventional courses; 84 middle- and upper-ability pupils in each of the control and experimental groups, matched for ability on the basis of predicted GCSE grades. No gender distribution provided for the sample.

Methods used to collect data

- Written questionnaire containing eight diagnostic questions
- Free response questions, exploring enjoyment of science and career intentions
- Data on projected pupil grades in forthcoming public examinations (General Certificate of Secondary Education, GCSE)

Data-collection instruments, including details of checks on reliability and validity

- Responses to eight diagnostic questions plus free response items
- Projected GCSE grades

Methods used to analyse data, including details of checks on reliability and validity

- Numerical rating of responses to diagnostic questions
- Statistical analysis (SD, standard error of the mean, critical ratio)
- Coding of responses to free response questions

Summary of results

- No significant differences were established in performance on diagnostic questions.

- Some key ideas were poorly understood as evident from written explanations, irrespective of approach.
- 20% of pupils indicated they wanted to study chemistry beyond the compulsory period, equally divided between each approach.
- The majority of those not pursuing their study further felt chemistry was not relevant for their career intentions.
- Pupils following the context-based course reported higher levels of interest in their science lessons than those following the more conventional course.
- 15% of the sample (32 students – 23 male, 9 female) intended to go on studying chemistry at A-level.

Conclusions

- '... this study suggests that there is little difference in promoting understanding of key chemical ideas between a context-based approach and a more traditional approach' (p 710)
- 'In terms of pupils' more general responses to science, their comments indicate that a context-based approach appears to be enjoyed by those following such a course and makes them more interested in what they are studying. None-the-less, it remains a concern that further study of chemistry beyond the compulsory period is not seen as a worthwhile route to take by the majority of pupils, irrespective of the course they follow, where choices are clearly made on he basis of career aspirations.' (p 710)
- Gender patterns for science subject choices formerly apparent at 14+ have now been transferred to 16+.

Weight of evidence A (trustworthiness in relation to study questions)

High: Large sample, with carefully matched intervention and control groups for predicted GCSE grades. Piloted and peer-reviewed questionnaire items. Statistical analysis of rated responses for significant differences between intervention and control groups.

Weight of evidence B (GA) (appropriateness of research design and analysis)

Low: This study is not an evaluation of gender effect.

Weight of evidence C (GA) (relevance of focus of review)

Low: Gender ratios in sample not provided; gender is not specific variable and a very small part of the study; questionnaire items on enjoyment and plans for future study appropriate for measuring attitude, not systematically reported; gender difference for career plans reflect narrow range of attitudes to science; situation partly representative (sample mainly average and upper-ability pupils).

Weight of evidence D (GA) (overall weight of evidence)

Medium low

Reid, N and Skryabina, E (2002) Attitudes towards physics

Country of study	UK: Scotland
Details of researchers	PhD student with supervisor from University of Glasgow
Name of programme	Standard Grade (context-based) and Higher Grade (traditional pre-university) school physics
Age of learners	The study covered primary to university students. However, the focus of this review is the context-based course taught at S3/S4 (15- to 16-year-olds) compared with the traditional course at S5/S6 (17- to 18-year-olds). Only details from that sub-set is included in this summary and in the review. The major part of the gender information is taken from Elena Skryabina's PhD thesis 'Students' attitudes to learning physics at school and university levels in Scotland', Glasgow University, 2000.
Type of study	Naturally-occurring evaluation

Aims of study

- What are the attitudes towards school science and physics courses in Scotland of 10- to 20-year-old pupils and students?
- What are the factors making physics at Higher Grade so popular a school subject choice?

Summary of study design, including details of sample

- Questionnaire survey, using a cross-age analysis
- No details of sample selection method. Actual samples: 103 (69 boys and 34 girls) in S3 and 152 (87 boys and 65 girls) in S4 from six schools, and 96 (68 boys and 28 girls) in S5/S6 from three secondary schools.
- Data from S3 and S4 were inspected separately. Data from S5 and S6 were combined as the work in those years was not considered to be qualitatively different.

Methods used to collect data

- Six-point Osgood scale (semantic differential format) questionnaires

Data-collection instruments, including details of checks on reliability and validity

- Osgood scales questions covered attitude to science lessons, self-perception, career intentions.
- No pilot
- Reliability of Osgood instrument was discussed as having been established by earlier authors. Validity of statements and methods was confirmed by appropriate independent, education professional.
- Authors acknowledged disadvantage of cross-age design: lack of control of population characteristics and risks of self-selection of those taking physics from S3 onwards.

Methods used to analyse data, including details of checks on reliability and validity

Quantitative:

- Scores from Osgood responses were aggregated per item in positive, neutral and negative categories, and then compared for age groups and for gender.
- Statistical differences calculated using χ^2 value, Spearman rank-order coefficient and Kendall's tau-b.

Summary of results

Only significant results given.

S3 and S4:

- Only one significant difference in attitudes of S3 boys and girls towards their physics lessons was found; lessons appeared to be more complicated for S3 girls than boys.
- No differences in attitudes to physics were found between boys and girls in S4.
- Comparison of S3 girls with S4 girls and S3 boys with S4 boys showed few significant differences. However, there was a growing enjoyment from lessons for girls and a growing of importance of Standard Grade physics lessons for boys as they move from S3 to S4.
- There were no differences between sexes in perception of self in physics lessons in S3.
- At S4, significantly more boys than girls thought 'Physics is definitely "my" subject'. Comparison of S3 with S4 responses revealed that girls were feeling positive about themselves at S4 in terms of intellectual growth and obtaining new skills. Boys at level 4 were found to be stronger than S3 boys in their perceptions about physics as 'definitely "my" subject' (p 75-76).
- 68% of boys and of girls in S3 indicated that they would continue with physics. This rose to 92% girls and 89% boys in S4.

S5/6:

- At S5/6 there was no difference in the attitude of girls and boys towards their physics lessons.
- There was only one gender difference with respect to pupils' self perception; significantly more girls than boys felt that they were growing intellectually in their Higher Grade physics course.

- 86% of girls and 80% of boys in S5/6 wished to study at university but only 13% of girls (3) and11% of boys (6) were planning to study physics.

S4 compared with S5/6:

- S4 boys were significantly more positive than S5/6 boys in respect of enjoyment of lessons, interest in lessons and perceived importance of lessons.
- S4 girls thought their lessons more enjoyable and more important than S5/6 girls.
- Both S5/6 boys and girls felt that they were obtaining significantly fewer new skills than their younger peers at Standard Grade.

Physics topics:

- S4 boys showed more interest than girls in a range of physics topics unrelated to the syllabus. There was no gender difference at S3 and S5/6. Girls showed a preference for 'Explanations of natural phenomenon' and 'Medical applications'. Boys showed a preference for 'Explanations of how technical objects function' and 'Personal benefit from knowledge'.
- Girls' interests in doing practical work declined significantly between S3 and S4. Boys demonstrated quite strong and stable interests and enjoyment from practical work up to S4 and then a significant decline at S5/6.
- A significant decline in boys' enjoyment in study of equipment-making was observed with age.

Conclusions

There are almost no differences in attitudes towards physics lessons between boys and girls doing S3 and S4 Standard physics courses with respect to interest towards lessons, enjoyment of lessons and importance of lessons.

Attitudes to self in physics lessons show a tendency for significant improvement with age for three questions for girls and one for boys.

Girls and boys show a high interest in continuing to study physics onto Higher Grade especially in S4.

Standard Grade physics, an applications-based approach, looks more interesting, enjoyable and important than the principles-based, more traditional Higher Grade course, especially for boys.

The principles–based approach of the Higher Grade physics course had some negative influences on boys', but not girls', attitudes towards self perception in physics: they find lessons less interesting, less important and less enjoyable.

The decline in positive attitude to science between S4 and S5/6, despite the self-selecting nature of the group (i.e. post-compulsory age), may reflect the importance of syllabus (S4 applications-based compared with S5/6 principles-based) type in promoting pupils' attitudes towards the subject.

The Standard Grade physics course shows good retention of girls into the Higher Grade physics.

The fall in interest in S4 girls may be a consequence of gender stereotypy as this is a decision-making stage.

Boys demonstrated a level of interest significantly higher than girls in topics related to technical objects and the way they function.

Girls preferred topics related to social, medical aspects of physics.

Both sexes demonstrated high levels of interest in topics related to explaining natural phenomena, practical benefit from knowledge and practical work.

Studying theory was found to be the least enjoyable activity for all pupils at every age group.

Weight of evidence A (trustworthiness in relation to study questions)

Medium high: Good samples (more than 95 at each age). No details of sampling method but acknowledgement that cross-sectional design does not allow for controls, and limits comparability because of self-selection in older samples. Good Osgood scales items for questionnaires. Appropriate statistical analysis.

Weight of evidence B (GA)

Medium: Good sample sizes across the three stages but no sampling frame; comparison for gender major part of the design; benchmark data comes from previous ages but is not same sample of pupils but difficult to see how else to benchmark; explicit discussion and checks for validity and reliability for data-collection and data analysis.

Weight of evidence C (GA) (relevance of focus of study to review)

High

Gender ratios not equal but sufficiently close; gender explicit independent variable; measures (Osgood scale) highly appropriate for testing attitude to school science; considerable breadth of attitude measured under three categories (attitude to subject, self-perception and career intentions); representative of classroom situation.

Weight of evidence D (GA) (overall weight of evidence)

Medium high

Smith, G and Matthews, P (2000) Science, technology and society in transition year: a pilot study

Country of study	Ireland
Details of researchers	Seemingly a project undertaken by a teacher and supervisor from Trinity College, Dublin
Name of programme	No name provided. Specially devised one-year STS course aimed at fostering personal interests, curiosity and positive attitudes in pupils. Pupils are encouraged to raise, pursue and resolve everyday questions related to science. An interdisciplinary approach uses a diversity of teaching stimuli (newspapers and videos) other than the textbook. SATIS and Science across Europe are being used as resource materials.
Age of learners	14- to 15-year-olds (Transition Year)
Type of study	Researcher-manipulated evaluation

Aims of study

- To establish the effectiveness of the intervention in improving pupils' attitudes to science and technology
- To establish if it would increase pupils' uptake of (higher level) Leaving Certificate sciences

Summary of study design, including details of sample

- Small-scale evaluative pre-post intervention experimental design (see p 109).
- No sampling method reported. Actual sample includes all 60 Transition Year pupils at one Community College; choice presumably determined by access. Pupils are typically from deprived background with an inclination towards early school leaving and low uptake of science at higher levels. There are 37 pupils in the experimental (STS physics and biology modules) group, and 23 pupils in the control group using a traditional course (i.e. as taught in previous transition years). Pupils allocated equally to experimental and control group for gender; no information about ability distribution.
- One of the researchers (Smith) taught the biology modules; another teacher taught the physics modules. Same content is covered in both experimental and traditional 10-week modules.

Methods used to collect data

- Two sets of pre- and post-intervention questionnaires for all 60 pupils were used as quantitative measures.
- Recorded individual interviews are employed as qualitative measures. 12 pupils of varying ability (eight from experimental, four from traditional group) were selected for interview, as validation of the perceptions emerging from the questionnaires (see p 110).

Data-collection instruments, including details of checks on reliability and validity

- Questionnaire A consisted of 25 questions (see Table 2) concerned with pupils' perceptions of school science and of science in general (latter not covered by items in examples). Responses were made on a five-point, Likert-type scale. Pupils were also asked whether or not they intended to choose a science subject at senior level, and, if so, which one and the reasons for their choice.

- Questionnaire B consisted of 21 questions (see Table 2) which focused on pupils' perceptions of the way sciences were and should be taught (latter not covered in the items). Responses were also categorised using a Likert-type scale. A number of the questions were modified from a previously used 'pupils' attitude to science' inventory which might, to some extent, enhance the validity of the instrument.
- No piloting or peer validation of other items reported. The use of questionnaires and interviews enhances the reliability of the data.
- Interview schedule prompts are provided (Table 3). No pilot is reported.

Methods used to analyse data, including details of checks on reliability and validity

- Data were analysed using SPSS. Responses for each item are classified in levels: positive, neutral and negative. Frequencies for each level per item are calculated for both groups.
- T-tests are used for determining significant differences of level frequencies for experimental and control groups, for the whole sample, female and male pupils, respectively.
- No details are given of reliability of validity checks on the data.

Summary of results

Note: No numerical data are provided, only descriptions including statements about significant differences. Many results are reported as views on science whereas only views on *school* science have been gathered.

- Pupils in the experimental group developed more positive perceptions of school science (significant positive change in 19/25 items) than those in the control group (significant positive change in 7/25 items, one negative). The same occurs for perceptions of science teachers.
- After instruction, female pupils in the experimental group developed markedly more positive perceptions of science than their female counterparts in the control group. This finding was also true of male pupils. Female pupils in the experimental group developed more positive perceptions of science teachers than those in the control group. This was also the case for male pupils.

Note: Subject choices are not mutually exclusive, but this has not been mentioned. No attention drawn to distinctly lower percentage pre-intervention interest in science career for experimental group, so post-intervention increase in interest brings them on a par with control group.

- Numbers of pupils wishing to opt for physics and biology at senior level doubled in the experimental group (from 5 to 12, from 14 to 26), but remained the same in the control group (from 8 to 7, from 12 to 12); see Table 4.
- Of the pupils in the experimental group opting to pursue a science, twice as many stated interest as a reason for their choice at post-instruction compared with pre-instruction (Table 5). For pupils in the control group opting for further study in the sciences, interest as a motivator decreased for all subjects from pre- to post-intervention. (Table 6 actually shows no difference.)

Conclusions

The researchers concluded as follows:

- Science presented in a social context, with opportunities to talk about moral and ethical issues, and a focus on people and their predicaments, is more attractive to girls than the traditional topic-focused approach.
- STS courses are more motivating for lower-ability and demotivated pupils than traditional science courses.
- STS courses encourage further study in the sciences (including biology and physics) by demotivated pupils compared with traditional courses, thus narrowing the gap that usually exists between male and female pupils, with STS seeming to be a promising reform in addressing gender imbalance.

Weight of evidence A (trustworthiness in relation to study questions)

Medium low: Small sample in one school, selected purposefully for demotivated learners. Good pre-post intervention design for experimental and control groups balanced for gender, not for achievement. Interviews for validating questionnaire responses. Some items modified from existing attitude scales, otherwise little effort to check reliability/validity of instruments. Items do not cover attitudes to science (as claimed) but to school science. Standard validity problems with aggregating Likert-scale responses. Standard statistical tests used, but no numerical data provided.

Weight of evidence B (GA) (appropriateness of research design and analysis)

Medium: Small sample size; good comparison for gender in design; good pre-post data-collection on attitude; some item adaptation from existing instrument provide validity, but items did not cover half of the research variables; standard statistics provides some reliability/validity of data analysis, but no actual data provided.

Weight of evidence C (GA) (relevance of focus of study to review)

Medium: Good balance of gender in experiment/control groups; gender explicit variable; large number of attitude items, but focus on school science and how teachers teach science, not on science in general and how science should be taught (as claimed); measures have reasonable range for attitude testing; prevalence of low achieving, de-motivated pupils in sample.

Weight of evidence D (GA) (overall weight of evidence)

Medium

Weight of evidence B (LA) (appropriateness of research design and analysis)

Medium Low: Small sample size; no comparison for ability in design; good pre-post data-collection on attitude; some item adaptation from existing instrument provide validity, but items did not cover half of the research variables; standard statistics provides some reliability/validity of data analysis, but no actual data provided.

Weight of evidence C (LA) (relevance of focus of study to review)

Medium: Detailed information on low ability/disaffected; low ability not an explicit variable; large number of attitude items, but focus on school science and how teachers teach science, not on science in general and how science should be taught (as claimed); measures have reasonable range for attitude testing; prevalence of low achieving, de-motivated pupils in sample.

Weight of evidence D (LA) (overall weight of evidence)

Medium low

Smith LA, Bitner BL (1993) Comparison of formal operations: students enrolled in ChemCom versus a traditional chemistry course

Country of study	USA
Details of researchers	One school-based and one university-based science education researcher
Name of programme	ChemCom, a year-long course built around eight societal issues related to chemistry aimed at lower achieving ‘non-science major’ pupils with less mathematics, covering the same basic chemistry concepts, basic vocabulary and laboratory skills as traditional courses. Each unit is laboratory-oriented and contains decision-making activities to apply chemistry to solve problems. Materials include pupils-centred, activity-based, issue-oriented activities to practise individual problem-solving and co-operative learning.
Age of learners	11–16
Type of study	Researcher-manipulated evaluation

Aims of study

To test three null hypotheses (N0):

- There is no significant difference in student gain in formal operational thinking in courses using an STS approach (ChemCom) and a traditional approach (GenChem) (H01).
- There is no significant difference in reasoning level of pupils in ChemCom and GenChem courses (H02).

- There is no significant difference in reasoning abilities between males and females enrolled in ChemCom and GenChem courses (N03).

Summary of study design, including details of sample

- Random selection of classrooms from five schools in Midwestern City
- 123 pupils. Experimental; group (ChemCom) with N=60, control (GenChem) with N=63 pupils, both equally split between boys and girls.
- The experimental group has non-science majors; control group has science-majors. Similar age distribution, all completed two semesters and did not repeat the courses.

Methods used to collect data

- Use of Piagetian test for formal reasoning level; abbreviated GALT (Group Assessment of Logical Thinking) paper-and-pencil test administered before and full GALT after interventions.

Data-collection instruments, including details of checks on reliability and validity

- GALT test contains MC + justification items, with reliability of 0.85 compared with Piagetian interview tasks. Four items each for testing conservation, proportionality, controlling variables, probability, correlational, combinatorial reasoning.
- Validity and reliability of GALT is well established.

Methods used to analyse data, including details of checks on reliability and validity

- Responses scored and aggregated to three reasoning levels: formal (score 8–12), transitional (5–7) and concrete (0–4).
- No inter-scorer agreement reported.
- Means and SD computed for pre- and post-tests for both groups; pre-existing differences compensated by analysis of covariance.
- χ^2 statistic used for analysing differences in pre- and post-test.

Summary of results

- H01 rejected: There is a significant difference in reasoning gain of pupils in ChemCom over those in GenChem.
- H02 accepted: There is no significant difference in reasoning levels between pupils enrolled in ChemCom and GenChem. The latter had better formal reasoning.
- H03 accepted: No significant difference existed between reasoning levels of males and females enrolled, and reasoning gains between males and females. In both cases, males had better reasoning and higher gains.

Conclusions

Difficult to align the findings and rejection/acceptance of the hypotheses with the conclusions quoted below:

- There was no significant difference in reasoning gain of pupils in ChemCom versus GenChem, indicating that the teaching strategy did not make a significant impact upon student gain in reasoning skills.
- A significant difference existed between pupils entering ChemCom and GenChem. (Yes, but not RQ)
- Significant differences existed between males and females enrolled in both ChemCom and GenChem, and reasoning gains between males and females. (Different in what direction?)

Weight of evidence A (trustworthiness in relation to study questions)

Medium low: Smallish sample, with pre-post high quality test (GALT) for control and experimental group with distinctly different entry level of reasoning levels. The small numbers (around 30) in each sub-sample, together with the major difference in entry level between experiment and control group, makes the rigorous statistics less applicable. Major confusion: the findings and conclusions seem to conflict. For the purposes of this review, we have taken the findings as conclusions.

Weight of evidence B (GU) (appropriateness of research design and analysis)

Medium high. Medium sample size, with clear sampling method; comparison for gender in design; good pre-post understanding data; reliability/validity for instrument good, except that prerequisite content knowledge was not explored; reliability/validity of analysis not clear.

Weight of evidence C (GU) (relevance of focus of study to review)

Medium high: Good balance gender ratios; gender explicit variable; measures good for testing science understanding; limited breadth of science understanding (reasoning only); representative situation.

Weight of evidence D (GU) (overall weight of evidence)

Medium

Sutman F, Bruce M (1992) Chemistry in the community – ChemCom

Country of study	USA
Details of researchers	Funder of curriculum development and researcher at Temple University Philadelphia
Name of programme	Chemistry in the Community (ChemCom), a year-long course build around eight societal issues related to chemistry, aimed at lower achieving 'non-science major' pupils with less mathematics, covering the same basic chemistry concepts, basic vocabulary and laboratory skills as traditional courses. Each unit is laboratory oriented and contains decision-making activities to apply chemistry to solve problems. Materials include pupils-centred, activity-based, issue oriented activities to practise individual problem-solving and co-operative learning.
Age of learners	15–17 (90% in Grades 10–12, some in college)
Type of study	Naturally-occurring evaluation

Aims of study

- (a) To evaluate the perceived effectiveness of ChemCom course in achieving its (11 listed) objectives
- (b) To survey student achievement at the end of ChemCom chapters and course
- (c) To determine the nature of student and teacher populations taking ChemCom
- (d) To assess teachers' responses to instructional practices advocated in ChemCom

Summary of study design, including details of sample

- Post-hoc survey of and pilot evaluation over five years of perceptions of five groups (Steering Committee members, writing team members, editors, teachers and pupils) of the pilot version of the intervention. The samples for the first three groups are assumed to cover the whole respective populations.
- Post-hoc experiment-control group comparison for pupils learning through ChemCom and through traditional courses.
- Actual samples of pupils (3,700) and teachers (84) were selected equally from rural/suburban and urban schools, split in an experimental (ChemCom) and a control (traditional) group. No details for experimental/control group selection/allocation methods.
- Teacher sample from seven diverse sites across the country is claimed to be 'realistic', with characteristics for experience and training provided.
- About 60% of student sample in pre-college programme, 40% in non-college programme, the vast majority in Grades 10–12.

Methods used to collect data

- For (a): Likert (five-scale) items questionnaire for all five groups assessing perceived effectiveness of course
- For (b): Multiple-choice, conceptual test for pupils, available in teachers guide

Data-collection instruments, including details of checks on reliability and validity

- Opinionnaire of 11 five-point Likert scale items solicited for views on each objective on the need for its inclusion in the course, and the degree of having been met. No detail on how items were piloted.
- Multiple-choice, conceptual test items (ratio conceptual:application=2:1) extensively validated by panel of five chemists/chemistry educators for coverage, content validity, readability, etc. Piloted with pupils following a traditional or no chemistry course. Availability of tests in teachers' guide may decrease validity of responses. Also validation through correlation of responses to application items with Longeot Test for Cognitive Development written by about 100 pupils (positive correlation of 0.4). Test reliability factor is 0.80 (unclear how calculated).

Methods used to analyse data, including details of checks on reliability and validity

- For Likert scale scores, the Provus Discrepancy Model was used to calculate the mean discrepancy value for each group (all 0.38-1.08). Standard problems about validity of aggregated Likert scores remain.
- Some significance testing was done for differences between experimental and control student groups, but no detail provided.

Summary of results

- All eleven objectives were considered valid (required and implemented) in the ChemCom course by all groups.
- Pupils completing the ChemCom course significantly outperformed pupils completing a traditional course on items designed to assess both chemistry content and the ability to apply this content knowledge.
- Teachers' ideas about the nature of high-school science conflict between the 'societal objectives' and the 'further training objectives' (pre-college analytical skills development).

Conclusions

- ChemCom (studied during development and field testing) is a highly effective intervention in securing higher achievement than traditional courses for lower-ability pupils.
- ChemCom does not deter pupils from learning chemistry concepts.
- ChemCom motivates more high-ability pupils than traditional courses to pursue higher level chemistry courses (no evidence provided).

Weight of Evidence A (trustworthiness in relation to study questions)

Medium Low: Use of total population of three groups, and large sample for teachers and pupils. Much of the paper is concerned with method (reliability/validity of data-collection) and the justification for the ChemCom programme. Less detail is given of analysis of results. Unclear how bias has been avoided in constructing experimental and control groups. No pre-intervention data, which is essential for this study as ChemCom is intended for lower-ability pupils. No information provided for data-collection and analysis to address the last research question.

Weight of evidence B (LU) (appropriateness of research design and analysis)

Medium low: RQ (b) is the only one relevant. Large sample but no sampling method; no comparison for low ability in control group; only post-data on achievement; solid checks on instrument validity; difficult to see validity of analysis method.

Weight of evidence C (LU) (relevance of focus of study to review)

Medium High: More non-science majors than low-ability learners (criterion for selection); low ability/non-science discrete element in intervention; highly scrutinised multiple-choice items appropriate for testing achievement; understanding of concept and application tested but limited reporting; highly representative for learners in classrooms.

Weight of evidence D (LU) (overall weight of evidence)

Medium

Wierstra RFA (1984) A study on classroom environment and on cognitive and affective outcomes of the PLON-curriculum

Country of study	The Netherlands
Details of researchers	Researcher at the University of Utrecht
Name of programme	Physics Curriculum Development Project (PLON)
Age of learners	15–16
Type of study	Researcher-manipulated evaluation, non-randomised controlled trial

Aims of study

- To investigate some aspects of inquiry-learning
- Three domains of the effects of the MAVO (intermediate level) of the PLON curriculum – cognitive achievement, affective and attitudinal outcomes, and the learning environment in the classroom – were measured.
- The purpose was to assess effects within each of these domains as well as the relationships between them.

Summary of study design, including details of sample

- Researcher-manipulated evalution
- A controlled intervention with data collected during and towards the end of the intervention
- Nine classes (experimental) were following the PLON curriculum and six the non-PLON curriculum (control). There was no prospective allocations but use of pre-existing differences to create the comparison groups.
- The sample was 297 boys and 101 girls in Dutch secondary schools.

Methods used to collect data

- Quantitative

Data-collection instruments, including details of checks on reliability and validity

- Individualised environment questionnaire using ten items on a five-point scale to assess pupil perceptions of actual and preferred learning environment
- Attitude (enjoyment/relevance/affect) questionnaire 12 items using a five-point Likert scale
- Physics achievement tests from PLON and traditional physics exams
- Established tests were used and checks were carried out for reliability and validity.

Methods used to analyse data, including details of checks on reliability and validity

- Correlations, t-tests and MANOVA
- Checks for reliability and validity of data analysis were not mentioned explicitly but standard statistical tests were used.

Summary of results

- In both the PLON group and the control group, it was found that enjoyment and relevance were correlated with the two achievement scores.
- With regard to the cognitive outcomes, no differences were found in the total test. However, PLON pupils did significantly better on test A (PLON based on PLON exams) and the control group did significantly better on test B (based on traditional physics exams).
- In PLON classes, there is considerably more enquiry-learning than in control classes (as reported by pupils).
- In both groups it was found that a learning environment with more emphasis on inquiry-learning has a positive influence on the attitude to physics; moreover, in the PLON group, there were significant effects of inquiry learning (as expressed as class scores) on cognitive achievement. The greater amount of inquiry learning has a more favourable affective outcome for the PLON group.
- Gender also had an impact. Boys showed higher scores on enjoyment and both achievement tests. Girls had a slightly more positive belief in learning from investigation.

- Pupils' preference for inquiry learning is strongly influenced by the actual amount of inquiry learning he or she experiences.

Conclusions

'In spite of the more positive attitudes of PLON pupils to physics and in spite of the positive correlations between attitude and achievement, there was no evidence for a better achievement by PLON pupils' (p 280).

'We found large and significant differences between PLON classes and control classes with regard to the amount of inquiry-learning as measured by the perceptions of pupils. However, there were also large differences amongst PLON classes in the amount of inquiry-learning. This is because of the PLON philosophy of giving teachers a great deal of freedom to create their own class environment' (p 280).

'Never-the-less, this study shows that the more inquiry based the classroom environment, the more favourable the pupils' attitude to physics (in both groups).'

Gender also had an impact. Boys showed significantly higher scores than girls on both achievement tests. Girls had a significantly slightly more positive belief in learning from investigation. Boys showed significantly higher scores on enjoyment.

Weight of evidence A (trustworthiness of study findings in relation to answering study question)

Medium: Sample size was good and control classes sampled at random were used. No pre-testing. Reliability and validity checks were carried out for data-collection and standard statistical tests used for analysis. Findings closely map onto research questions but there are gaps in evidence-reporting, such as on recruitment and data: only correlation between various tests are reported rather than data from various tests.

Weight of evidence B (GA) (appropriateness of research design and analysis for addressing review question)

Medium: Large sample and sampling frame used; comparison for gender in the findings not explicit in the design; no pre-intervention data; attitude was measured using previously established tests (five-point Likert with reliability test alpha coefficient) but did not present the data only the statistical comparisons; used standard statistical test (univariate tests) for analysis of gender and attitude (enjoyment).

Weight of evidence C (GA) (relevance of focus of study for addressing review question)

Medium high: Balance was not ideal (three boys to one girl); gender was one variable (type of curriculum and teacher were the others); no details of measures used for testing attitude but were used in previous study; breadth of measures was good, highly representative of learners in classroom.

Weight of evidence D (GA) (overall weight of evidence)

Medium

Weight of evidence B (GU) (appropriateness of research design and analysis for addressing review question)

Medium: Large sample and sampling frame used; comparison for gender in the findings not explicit in the design; no pre-intervention data; understanding was measured using previously established tests and checks for reliability between the PLON and traditional test questions, but did not present data-only statistical analysis; used standard statistical tests for analysis of understanding (MANOVA).

Weight of evidence C (GU) (relevance of focus of study for addressing review question)

Medium high: Balance was not ideal (three boys to one girl); gender was one variable (type of curriculum and teacher were the others); measures were appropriate for testing understanding (tests from each curriculum); breadth was broad (18 items from each curriculum); highly representative of learners in classroom.

Weight of evidence D (GU) (overall weight of evidence)

Medium

Wierstra RFA, Wubbels T (1994) Student perception and appraisal of the learning environment: core concepts in the evaluation of the PLON physics curriculum

Country of study	The Netherlands
Details of researchers	Researchers at the University of Utrecht
Name of programme	Physics Curriculum Development Project (PLON). The study was based on a four-week teaching unit ‘Traffic’ which encompassed the teaching and learning of mechanics.
Age of learners	15–16
Type of study	Researcher-manipulated evaluation

Aims of study

- To discover to what extent the PLON curriculum and, in particular, its reality- and activity-centeredness result in instruction that pupils experience as meaningful
- To find out how these curriculum characteristics affect affective and cognitive pupil outcomes

Summary of study design, including details of sample

- Non-randomised controlled trial
- Pupils in general, mixed sex, secondary education in schools in the Netherlands in 1988. The experimental group was made up of eight classes (totalling 209 pupils) taking the PLON physics curriculum and the control group made up of 16 classes (with 355 pupils) taking two traditional physics curricula (Middelink and Scoop).

Data-collection instruments, including details of checks on reliability and validity

- Pupils cognitive outcomes data were collected by a 19-item conventional, multiple-choice, standard physics tests covering the knowledge and performance common to the PLON and control curricula.
- Data on affective outcomes (pupil appreciation of the unit, perceived instructiveness, general attitude to physics and pupil entering attitudes) were measured with Likert-type questionnaires.
- Perceptions of the classroom environment were measured by a questionnaire on a five-point scale.
- Details of pupil gender were also taken into account when scoring for cognitive outcomes.
- Considerabale discussion of the reliability and validity of the data-collection tools.

Methods used to analyse data, including details of checks on reliability and validity

- Analyses of variance
- Correlation analyses between various aspects measured were also carried out to check for validity and reliability.

Summary of results

- PLON pupils experienced the lessons as significantly more reality- and activity-centred than the control pupils.
- Scores for PLON pupil appreciation and perceived instructiveness of mechanics lessons were significantly less than those for pupils on traditional courses. This was in contrast to expectations.
- The influence of reality-centredness on the pupils’ perceived instructiveness seems to be stronger than the influence of activity-centredness.
- There was no significant difference between the two groups with respect to cognitive outcomes.
- A significant difference in cognitive outcomes/understanding is shown in Table 9 (p 451) in the ANOVA but not commented on. As the mean differences are not reported, it is not clear what those differences are.
- No significant gender differences for appreciation of mechanics were found (ANOVA Table 7).

Conclusions in relation to understanding

‘There appears to be no significant effect of the curriculum on the [cognitive] test scores. Unlike some critics’ expectations, the PLON curriculum’s emphasis on physics in everyday life was not paid for with lower-standard cognitive outcomes’ (p 451).

A difference between the sexes was found in understanding (but no details given of actual mean scores or of what that difference was).

Conclusion in relations to attitude

'Our results in the evaluation of the unit Traffic indicate that reality-centeredness of the learning environment is a notable characteristic of the learning environment promoting perceived instructiveness. … Activity-centeredness also seems to promote affective outcomes, but more the appreciation of lessons than the perceived instructiveness. …In the study about Traffic we found that that the learning environment was perceived as more reality- and activity-centred in the PLON classes than in the control classes. Despite this difference and the correlations between reality- and activity-centredness and the affective outcomes, these outcomes tended to be lower in the PLON tan the control classes' (p452).

No gender differences in attitude were found.

Weight of evidence A (trustworthiness of study findings in relation to answering study question)

High: Four explicit research questions; control used; large sample of 24 classes, gender considered in the design; prior physics grade registered; pre-test of attitude to physics generally and also to types of instruction given; reliability tests carried out; validity derived from previously published questionnaire for affective outcomes. Results presented in context of research questions. However most of the focus is on the curricula rather than gender.

Weight of evidence B (GA) (appropriateness of research design and analysis for addressing review question)

Low: This is not an evaluation of gender. Gender is considered only as a covariable to be accounted for in the statistical analysis. It is not discussed in the results section or the conclusion.

Weight of evidence C (GA) (relevance of focus of study for addressing review question)

Medium low: Balance not given but boys and girls included; gender is wrapped up in the intervention; little detail given of attitude measures; little detail given of breadth of measures; little detail given of situation.

Weight of evidence D (GA) (overall weight of evidence)

Medium low

Weight of evidence B (GU) (appropriateness of research design and analysis for addressing review question)

Low: This is not an evaluation of gender. Gender is considered only as a covariable to be accounted for in the statistical analysis. It is not discussed in the results section or the conclusion.

Weight of evidence C (GU) (relevance of focus of study for addressing review question)

Medium low: Balance not given but boys and girls included; gender is wrapped up in the intervention; little detail given of understanding measures; 19 questions used for understanding; little detail given of situation.

Weight of evidence D (GU) (overall weight of evidence)

Medium low

Yager, RE and Weld, JD (1999) Scope, sequence and co-ordination: The Iowa Project, a national reform effort in the USA

Country of study	USA
Details of researchers	Academic researchers from the universities of Iowa and Oklahoma State.
Name of programme	Scope, Sequence & Co-ordination (SS&C): The Iowa Project. This STS approach used a constructivist learning model in line with the National Science Education Standards (USA). Compared with traditional courses, the outcomes of SS&C include (1) a more scientifically literate citizenry; (2) increased science uptake at advanced level, especially by females and minorities; (3) greater understanding of science content; (4) new approaches to the use of learning materials; and (5) more effective forms of student learning. Basic elements include a revisiting of science concepts in different contexts, experience with phenomena before introduction of terminology, assessment of depth of understanding rather than facts, and application of science to resolve personal and societal problems.
Age of learners	12–14
Type of study	Researcher-manipulated evaluation: non-randomised controlled trial

Aims of study

Hypotheses tested by authors (the review will focus only on the last two, student-related hypotheses):

- SS&C teachers have increased confidence to teach science and increased understanding of the nature of science and technology than control group teachers.
- SS&C teachers can use constructivist teaching strategies and action research projects better than control group teachers.
- Pupils achieve significantly better in SS&C courses than they do in traditional courses in six content areas: concept, process, application, creativity, attitude and worldview (descriptors of each provided).
- Low-ability pupils and female pupils are especially served by SS&C science in the same six domains.

Summary of study design, including details of sample

- Pre-post intervention design for equivalent experimental and control groups of teachers and pupils.
- The main student sample consisted of 5,270 pupils (2,557 female, 2,673 male) in the experimental group, and 1,320 (632 female, 688 male) in the control group. In both cases, these groups were roughly equally distributed over Grades 6-8. Pupils came from classes from all teachers involved in the SS&C intervention in five Iowa districts. The sample included at least two parallel classes taught by each teacher: in some, using the SS&C approach (experimental group); in others, the traditional approach (control group). Random allocation of classes to treatments was done by school counsellors, and care was taken that these classes were representative of the demographic composition of the Iowa districts involved.
- A subset of 364 experimental (187 female, 177 male) and 359 control pupils (180 female, 179 male) were involved in the career-orientation comparison. No indication how these were selected.

Methods used to collect data

- Concept-specific tests to measure student achievement for understanding of concepts and applications
- General questionnaires for measuring student views of process, creativity, attitude and worldview

Data-collection instruments, including details of checks on reliability and validity

- Concept and application tests were topic and unit tests provided by national panels.
- General instruments were taken from the Iowa Assessment Package. Authors note that this may favour the non-SS&C group.
- Career-orientation was determined by a four-point Likert scale questionnaire of five items.
- Content validity of all tests was checked by peer validation.

- Test-retest procedures within one week were used to establish reliability of all tests: consistent agreement range of 0.88–0.92.

Methods used to analyse data, including details of checks on reliability and validity

- Concept and application test scores were transferred into individual ‘improvement rates’ to allow for comparison (t-tests) across different topics (Tables 11 and 12).
- Post-intervention responses for pupils’ views on process, creativity, attitude and worldviews were compared for experimental and control groups (using ANCOVA) with pre-intervention scores as covariate (Tables 13–16). The same was done for comparisons of career-orientation between experimental and control groups (Table 2).
- Post-intervention responses for girls’ and boys’ views on process, creativity, attitude and worldviews were compared for experimental and control groups (using ANCOVA) with pre-intervention scores as covariate (Table 18). The same was done for comparisons of career-orientation between girls and boys in the experimental and control group respectively (Tables 3 and 4).
- Post-intervention responses for low- and high-ability pupils’ views on process, creativity, attitude and worldviews were compared for experimental and control groups (using ANCOVA) with pre-intervention scores as covariate (Table 17).
- The analysis uses as a main comparison the pupils of the same teacher in experimental and control groups, thus increasing validity.

Summary of results

- Pupils achieve better results in SS&C courses than in traditional text-based courses in all six achievement areas: concept, application, process, creativity, attitude and worldview.
- Pupils in SS&C classes show a significantly increased positive attitude to science as a career (four out of five items), whereas non-SS&C pupils only showed a limited improvement in career-orientation (one out of five items).
- Both female and male pupils in SS&C classes improved their learning in concept, application, process, creativity, attitude and worldview significantly more than non-SS&C pupils.
- Female pupils in SS&C classes improved their attitudes toward science and their understanding of the nature of science significantly more than male pupils in these classes.
- Both high- (upper one-third) and low- (lower one-third) ability pupils in SS&C classes improved their learning in application, process, creativity, attitude and worldview significantly more than non-SS&C pupils.
- Low-ability pupils in SS&C classes improved their learning in concept, application, process and attitude significantly more than high-ability pupils in these classes.

Conclusions

As above

Weight of evidence A (trustworthiness in relation to study questions)

High: Very large sample, carefully selected to represent the Iowa school population, for socio-economic, race, ESL variables. Controls used with pre-post testing. Established instruments well tested for reliability and validity. Careful administration of the tests, and detail for the statistical analysis, data from different teachers analysed separately.

Weight of evidence B (GA) (appropriateness of research design and analysis)

High: Large samples for gender and appropriate sampling frame; good comparison for gender in design; benchmark data on attitude included; solid checks on reliability/validity for data-collection (test-retest reliability of 0.88–0.92, non-teacher administration, use of previously validated items); solid checks on reliability/validity for data analysis for gender and attitude (ANOVA statistics, T-tests). Well reported.

Weight of evidence C (GA) (relevance of focus of study to review)

High: Balanced gender ratio in design; gender is explicit independent variable; measures not provided, but from well-validated Iowa tests; breadth of measures covers two of six domains, with some detail for career choices; situation representative of classrooms.

Weight of evidence D (GA) (overall weight of evidence)

High

Weight of evidence B (GU) (appropriateness of research design and analysis)

High: Large samples for gender and appropriate sampling frame; good comparison for gender in design; benchmark data on understanding included; solid checks on reliability/validity for data-collection (test-retest reliability of 0.88–0.92, non-teacher administration, use of previously validated items); solid checks on reliability/validity for data analysis for gender and understanding (t-tests for improvement rate). Well reported.

Weight of evidence C (GU) (relevance of focus of study to review)

High: Balanced gender ratio in design; gender explicit variable; measures not provided, but from well-validated Iowa tests; breadth of measures according to teaching topics; situation representative of classrooms.

Weight of evidence D (GU) (overall weight of evidence)

High

Weight of evidence B (LA) (appropriateness of research design and analysis)

High: Large samples for low ability, sampling method relies on ability grouping within the sample; good comparison for ability only in findings; benchmark data on attitude included; solid checks on reliability/validity for data-collection (test-retest reliability of 0.88–0.92, non-teacher administration, use of previously validated items); solid checks on reliability/validity for data analysis for ability and attitude (ANOVA statistics, t-tests).

Weight of evidence C (LA) (relevance of focus of study to review)

Medium: Lower one-third probably representative for low ability; low ability wrapped up in the intervention; measures not provided, but from well validated Iowa test; breadth of measures unclear, but from well validated Iowa tests; situation representative of classrooms.

Weight of evidence D (LA) (overall weight of evidence)

Medium high

Weight of evidence B (LU) (appropriateness of research design and analysis)

High: Large samples for low ability, sampling method relies on ability grouping within the sample; good comparison for ability only in findings; benchmark data on understanding included; solid checks on reliability/validity for data-collection (test-retest reliability of 0.88–0.92, non-teacher administration, use of previously validated items); solid checks on reliability/validity for data analysis for ability and understanding (t-tests on improvement rates).

Weight of evidence C (LU) (relevance of focus of study to review)

Medium high: lower one-third probably representative for low ability; low ability wrapped up in the intervention; measures from validated Iowa test, but not provided; breadth of measures assured as according to topics taught; situation representative of classrooms.

Weight of evidence D (LU) (overall weight of evidence)

High

Zoller *et al.* (1990) Goal attainment in science-technology-society (STS) education and reality: the case of British Columbia

Country of study	Canada: British Columbia
Details of researchers	Researchers from the universities of Haifa (Israel) and British Columbia (BC) and from RECSAM, Malaysia
Name of programme	British Columbia Junior Secondary Science Program (BC JSSP), also ST 11 Intended to provide opportunities to (i) develop an appreciation of the interactive nature of S-T-S; (ii) gain knowledge of technologies as applications of science; and (iii) develop the ability to respond critically to technological issues. The target group of the course contains non-science majors. The course consists of 16 modules, of which four are compulsory, and a further four are selected from the remaining twelve.
Age of learners	15–17; Grade 11
Type of study	Researcher-manipulated evaluation

Aims of study

To establish goal attainment of STS courses implemented on a large scale:

1. What has been the impact of ST1 teaching on pupils' views of the relationship S-T-S?
2. To what extent have the following five sub-goals (out of a total of 13) been attained during year I of implementation?
    - Understanding that society controls technological development (sub-goal 1.02)
    - Understanding that society influences and responds to scientific activity (sub-goal 1.03)
    - Understanding that technology is both a cause and result of scientific activity (sub-goal 1.04)
    - Understanding that technology is an application of the concepts and principles of science (subgoal 2.01)
    - Recognising that decisions concerning scientific and technological issues are influenced by values (sub-goal 3.01)

Summary of study design, including details of sample

- *Ex post facto* evaluation of pilot implementation. Selected five sub-goals considered indicators for implementation effectiveness of total programme (no attempt to cover other eight sub-goals).
- Six public/comprehensive schools within the greater Vancouver area were selected with 'slight adjustment' for non-specified stratification. A questionnaire was administered to a sample (N=473) of 12th graders: 101 (41 males, 60 females) in experimental group who completed the ST 11 course, 37% sample in randomly selected classes from 274 pupils (gender balanced) in the population of ST 11 pupils in the six schools.
- The control group contained 276 pupils (130 males, 146 females), a 19% sample in randomly selected classes from the 1,496 non-ST 11 pupils in the six schools. Five of six schools classified the population as non-academic.

Methods used to collect data

Questionnaire with four VOSTS items completed after the courses

Data-collection instruments, including details of checks on reliability and validity

- Questionnaire consisted of four pertinent statements of the VOSTS inventory form CDN testing respondents' beliefs regarding the five sub-goals. (Previous testing of validity and reliability by Aikenhead *et al.*, 1987. Note that although the reliability and validity checks say 'any combination of items may be used efficiently and reliably', no justification is given for the very limited number of items used.) Four items are provided *verbatim*.
- No pre-intervention data were collected, although some data (without detail) of a cohort in the early stage of their ST11 course are used to argue that the pre-intervention views of both groups are likely to be comparable.

Methods used to analyse data, including details of checks on reliability and validity

- Responses for experimental and control group were used to construct three or four 'response profiles' for each of the four items for both groups through clustering similar responses, based on the researchers' judgements.
- Similar groups were constructed for response profiles for male and female groups within experimental and control groups
- Significant differences between response profiles within each item are calculated using χ^2 tests
- No mention of independent clustering of responses (and thus validity check) when developing response profiles

Summary of results

- For three out of the four items, ST11 pupils responded significantly differently.
- ST11 pupils (also within the samples of male and female pupils) favoured the public (against scientists and engineers) deciding on issues of world food production. However, most pupils in both groups favoured shared responsibility, particularly the girls.
- ST11 pupils (also within the samples of male and female pupils) considered scientists to be responsible for the consequences of their discoveries, not only for the quality of the science *per se*. Pupils (both males and females) in both groups had polarised views here.
- ST11 pupils (males more than females) believed that personal opinion and moral values play a role in interpretation of facts, whereas non-ST11 pupils (males in particular) saw scientists as neutral and objective.
- Pupils in both groups (for both genders) failed to differentiate between the roles of science and technology.

Conclusions

- The ST 11 intervention has an impact on STS views of high-school pupils.
- The ST 11 achieved three of the five sub-goals: that is, the understanding that society controls technological development (sub-goal 1.02), and influences and responds to scientific activity (sub-goal 1.03), and the recognition that decisions concerning scientific/technological issues are influenced by values (sub-goal 3.01).
- The ST 11 course did not achieve sub-goal 2.01 (i.e. understanding the difference between science and technology).
- There is no indication about achieving sub-goal 1.04 (i.e. understanding that technology is a cause and a result of scientific activity).

Weight of evidence A (trustworthiness in relation to study questions)

Medium low: No pre-intervention data collected, but large sample and careful sampling method used. Reliable and valid instrument (VOSTS items), although only very few items included. Correspondence between outcomes tested and items included sometimes tenuous. Analysis method has medium reliability/validity and no cross-item analysis attempted.

Weight of evidence B (GA) (appropriateness of research design and analysis)

Medium: Large sample size; comparison for gender included in design; no benchmark data for gender; good items but very few, and sometimes not covering variables of RQs; medium data analysis.

Weight of evidence C (GA) (relevance of focus of study to review)

Medium: Gender ratios in experimental group slightly unbalanced; gender explicit independent variable; profiles provide good information about attitude to science but very few included; the breadth of attitude to science is limited to STS relationships; learners mainly 'non-academic' usually taught by non-science teachers.

Weight of evidence D (GA) (overall weight of evidence)

Medium