REVIEW

January 2005

A systematic review of the use of small-group discussions in science teaching with students aged 11-18, and the effect of different stimuli (print materials, practical work, ICT, video/film) on students' understanding of evidence

Review conducted by the Science Review Group

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

Name of group and institutional location

EPPI Review Group for Science
Department of Educational Studies, University of York, UK

Contact details

Dr Sylvia Hogarth
Department of Educational Studies
University of York
York YO10 5DD

Tel: 01904 433441
Email: sdh6@york.ac.uk

AUTHORS AND REVIEW TEAM

Dr Sylvia Hogarth	Department of Educational Studies, University of York, UK
Dr Judith Bennett	Department of Educational Studies, University of York, UK
Dr Bob Campbell	Department of Educational Studies, University of York, UK
Fred Lubben	Department of Educational Studies, University of York, UK
Alison Robinson	Department of Educational Studies, University of York, UK

REVIEW GROUP MEMBERSHIP

Dr Judith Bennett	Department of Educational Studies, University of York, UK (Review Group Co-ordinator)
Martin Braund	Department of Educational Studies, University of York, UK
Dr Bob Campbell	Department of Educational Studies, University of York, UK
Nick Daws	Ofsted Inspector and Science Inspector for Staffordshire LEA, UK
Steve Dickens	Head of Physics, Dixons City Technology College, Bradford, UK
Alison Fletcher	Head of Science, Huntington School, York, UK
Nichola Harper	Head of Chemistry, Aldridge School, Walsall, UK
Dr Sylvia Hogarth	Department of Educational Studies, University of York, UK (Review Group Research Fellow)
Professor John Holman	Department of Chemistry, University of York, UK, and Director of the Science Curriculum Centre, University of York

Declan Kennedy	University College, Cork, Ireland, and science textbook author
Dr Ralph Levinson	Institute of Education, University of London, UK
Fred Lubben	Department of Educational Studies, University of York, UK (Review Group Research Fellow)
Alyson Middlemass	Assistant Head Teacher and Head of Science, Elizabethan High School, Retford, UK
Professor Robin Millar	Department of Educational Studies, University of York, UK
Christine Prior	University of York, UK and Director of *Salters Advanced Chemistry*
Dr Mary Ratcliffe	University of Southampton, UK
Alison Robinson	Department of Educational Studies, University of York, UK (Review Group Information Officer and Administrator)
Daniel Sandforth-Smith	Institute of Physics (IoP), UK
Carole Torgerson	Department of Educational Studies, University of York, UK and member of the EPPI Review Group for English

ACKNOWLEDGEMENTS

The EPPI Review Group (RG) for Science is based in the Department of Educational Studies at the University of York, UK.

The EPPI RG for Science acknowledges financial support from the Department for Education and Skills (DfES) via the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre), and from the Department of Educational Studies at the University of York through direct financial support and through the use of the Higher Education Funding Council for England (HEFCE)-funded time for two of the core team members, Judith Bennett and Bob Campbell.

There are no conflicts of interest for the core team of RG members. Other members of the RG (John Holman, Robin Millar) are involved in the development of *21st Century Science*, a course currently in its pilot phase which will be advocating the use of small-group discussions. A number of members of the RG (Judith Bennett, Bob Campbell, John Holman, Robin Millar) were involved in the development of the *Salters* courses (*Science: the Salters Approach, Science Focus, Salters Advanced Chemistry, Salters Horners Advanced Physics*), all of which advocated the use of small-group discussions as one of a range of student-centred approaches in teaching.

LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
ASE	Association for Science Education
BEI	British Education Index
CBI	Computer-Based Instruction
CLE	Computer-Supported Learning Environment
DfEE	Department for Education and Employment
DfES	Department for Education and Science
EPPI-Centre	Evidence for Policy and Practice Information and Co-ordinating Centre
ERIC	Educational Resources Information Centre
GALT	Group Assessment of Logical Thinking
GCSE	General Certificate for Secondary Education
HPD-LC	Hypothetico predictive discussion learning cycle
HSD	Honest significant differences
ILL	Inter-library loan
LC	Learning cycle
LEA	Local education authority
MANOVA	Multi-analysis of variance
Ofsted	Office for Standards in Education
PGCE	Postgraduate Certificate in Education
POLS	Perspectives on learning science
PSE	Personal and social education
QCA	Qualifications and Curriculum Authority
RCT	Randomised controlled trial
SGD	Small-group discussion
SP	Seminal papers
SSCI	Social Sciences Citation Index
UK	United Kingdom
USA	United States of America
UYSEG	University of York Science Education Group
VOSTS	Views On Science Technology and Society

This report should be cited as: Hogarth S, Bennett J, Campbell B, Lubben F, Robinson A (2005) A systematic review of the use of small-group discussions in science teaching with students aged 11-18, and the effect of different stimuli (print materials, practical work, ICT, video/film) on students' understanding of evidence.
In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

The opinions expressed in this review are not necessarily those of the EPPI-Centre or the review funders. Responsibility for the views expressed remains solely with the review authors.

TABLE OF CONTENTS

SUMMARY

Background

This review builds on the work of an earlier systematic review (Bennett *et al.*, 2004) by continuing to focus on aspects of small-group discussions in science teaching. Small-group discussions have been strongly advocated as an important teaching approach in school science for a number of years, partly arising from a more general movement towards student-centred learning, and partly as a means of drawing on recommendations from social constructivist research, where it is seen as very important to provide students with an opportunity to articulate and reflect on their own ideas about scientific phenomena.

Several factors have come together recently to contribute to the current high levels of interest. These include the following:

- moves towards making changes in the school science curricula of a number of countries such that courses have an increased emphasis on the development of *scientific literacy*
- the most recent version of the National Curriculum for Science in England and Wales requiring that school students be explicitly taught about *ideas and evidence*
- the current interest in formative assessment as a key aspect of teaching
- a more general drive to improve students' *literacy skills*, formalised into the National Literacy Strategy in England and Wales (Department for Education and Employment (DfEE), 1998)

The systematic mapping of the area undertaken in the initial review revealed a wide range of relevant studies and facilitated the potential to explore a number of different aspects of the use of small-group discussion work in science teaching.

Aims

The review has two principal aims:

- to identify the ways in which small-group discussions are currently used in science lessons
- to look at the effects of small-group discussions on students' understanding of science ideas and attitudes to science

Review questions

The main review research question is as follows:

How are small-group discussions used in science teaching with students aged 11-18, and what are their effects on students' understanding in science or attitude to science?

The term 'understanding' encompasses science concepts, ideas about the nature of science and the methods of science. The term 'attitude' includes attitude towards science, attitude towards school science, motivation to learn, interest in science activities and career intentions.

The earlier review by the Review Group (Bennett *et al.*, 2004) led to a systematic map of research activity in the area and an in-depth review of studies addressing the question: *What is the evidence from evaluative studies of the effect of small-group discussions on students' understanding of evidence in science?*

Based on an update of the systematic map, the review reported here includes an in-depth review of studies addressing the question:

What is the evidence from evaluative studies of the effect of using different stimuli (print materials, practical work, ICT, video/film) in small-group discussions on students' understanding of evidence in science?

This particular research question was chosen because little research has been carried out into this aspect of small-group discussions and because there has been no attempt to put the evidence together.

For the purposes of this review, 'understanding of evidence' was defined as the understanding 'related to the collection, validation, representation and interpretation of evidence' (Gott and Duggan, 1996, p 793); that is, the ability to co-ordinate observations (primary or secondary data) with theory (models or concepts).

Methods

The review methods used are those developed by the Evidence for Policy and Practice Information and Co-ordinating Centre (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 in, or excluded from 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 combining the weighted findings of the collective studies

Results

The studies identified through the searching and screening processes established that three different stimuli (printed materials, computers and practical work) were being used in diverse ways in promoting discussions about scientific evidence with small groups of students. Aspects of the use of scientific evidence being investigated also varied: examples include seeking relevant information, identifying gaps in knowledge, evaluating, predicting, hypothesising, recognising anomalies, and testing, formulating and revising models or hypotheses. Similarly, measurements of the nature of the discussion or argumentation differed and included challenging, opposing, justifying, explaining, conceding and agreeing. Some papers also included measures of the metacognitive development of students about their understanding of the use of scientific evidence.

Ninety-four 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 small-group discussions that has been published in the English language, as summarised below.

- The majority of the studies report work that has taken place in the USA, the UK and Canada, although studies from many (13) other countries were included.

- Small-group discussions were used with all ages of student in the secondary age range.

- Most studies were carried out with mixed ability and mixed gender classes.

- The majority of work focused on small-group discussions in relation to students' understanding.

- The most common stimuli used to promote discussion were prepared curriculum materials, followed by practical work and then computer software.

- A diversity of measures was used to assess effects on understanding and attitude.

- Very little research has been done on small-group discussions in relation to the teaching of chemistry.

- Typical small-group discussions involved groups of three to four students emerging from friendship ties; they had a duration of at least 30 minutes.

- Typical small-group discussions had individual sense-making as their main aim (as opposed to, for example, leading to a group presentation).

- The most common research strategy was that of case study.

- Twenty-eight studies had experimental designs, of which 12 are RCTs.

- The most popular techniques for gathering data were observation, videotapes and audiotapes of discussions, interviews, questionnaires and test results.

Ten studies were included in the in-depth review which focused on the effect of using different stimuli on students' understanding of the use of evidence in small-group discussions. Following the application of the EPPI-Centre methods for assessing studies, seven were considered to be appropriately focused and of sufficient standard to form the basis for a synthesis of their findings.

The foci of the studies considered for the synthesis of this in-depth review vary somewhat. Therefore many of the findings have, on purpose, been cast in tentative terms because of their narrow evidence base.

The two findings that emerged most strongly from this review are as follows:

- Small-group discussion, focused on understanding the use of evidence, regardless of the prompt stimulus, is enhanced and focused by giving students some form of guidance on how to use that stimulus effectively. This guidance can be prior training in argumentation that provides instruction on how to use evidence or can be built into the structure or sequence of stimulus-based task.

- Second, a successful stimulus for students working in small groups to enhance their understanding of evidence has two elements. One requires students to generate their individual prediction, model or hypothesis which they then debate in their small group (internally driven conflict or debate). The second element requires them to test, compare, revise or develop that jointly with further data provided (externally driven conflict or debate).

Other findings of interest are given below:

- Prior knowledge can sometimes limit the understanding of evidence and its function. This can, for example, be the use of incorrect or inadequate factual knowledge or an idiosyncratic or inconsistent use of evidence to develop a hypothesis or test a model.

- Rich stimuli, such as those that provide complex and open-ended engagement, enhance opportunities for developing understanding of evidence.

Conclusions

Strengths of the review

The review has a number of strengths:

- The review focus is highly topical. The Review Group has already been contacted by potential users interested in the findings. Further evidence of the topicality comes from the range of countries in which studies have been undertaken and from the dramatic rise in relevant published papers since 1992.

- The review has served to establish that there is consistency in the research approaches that those working in the area feel are appropriate to researching practice related to the use of small-group discussions. Such approaches make use of quantitative data, but also draw extensively on qualitative data in the form of students' written responses, interviews and audiotapes of dialogue during discussions.

- End-users of the review findings have been closely involved at all stages of the review.

- Quality-assurance results are 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 stimulus as a discrete independent variable, which resulted in very little work emerging which related specifically to the in-depth review question. Of the ten studies that had an overall weighting of medium-high or medium, only in seven was the stimulus the variable that was being evaluated. As a result, only these seven studies were judged to be of reasonable weight with respect to the review question.

- 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. For example, there is considerable variety in the nature and purpose of the discussion tasks, in the data collected, and in the interpretation of the terms ‘evidence’ and ‘understanding of evidence’. Thus, teasing out the findings which specifically relate to small-group discussions and to particular stimuli was not easy, and 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.

Implications for policy

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

The review has *not* yielded any evidence that inclusion of any specific stimulus for small-group discussions adversely affects students’ understanding of the nature of evidence. However, it should also be noted that there is a scarcity of high quality research evidence in the area on which the in-depth review focused.

Where small-group discussions are advocated as a teaching approach, it is important to support this with guidance on running such discussions in a way which will increase the effectiveness of students’ learning. Such guidance should include advice to students on how to use materials for the purposes of discussion, as well as the stimulus materials themselves.

Implications for practice

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

The review has indicated that there is a diversity of ways in which the term ‘understanding of evidence’ is being interpreted. One implication for practice is therefore that teachers should be aware of this lack of clarity.

A further implication is that the success of small-group discussion, whatever the stimulus, depends in part on the students receiving some guidance on how to carry out or structure their discussions. That guidance might be written instruction, cues built into computer software or verbal support from teachers.

Presenting a task that offers opportunities for students to generated both their own input (e.g. own predictions, hypotheses, internal debate/conflict) and requirement to use that in conjunction with the stimulus provided by the teacher whether written, computer software or practical work (external debate/conflict) can be beneficial.

Tasks which are rich (i.e. complex and open-ended) are more likely to promote discussion and understanding of evidence in science than are simple or closed tasks.

Students' lack of sufficient factual knowledge of the subject of the task and/or of a systematic and consistent approach to the use of evidence can impede learning about evidence, unless support is given.

Teachers should also be aware of the lack of high quality research evidence in the area on which the in-depth review focused.

Implications for research

Secondary research

Exploration of additional areas of the systematic map would appear to be particularly helpful to provide a broader picture of research findings on small-group discussion work. Such areas would include the following:

- the use of small-group discussions in relation to the development of understanding of socio-scientific issues
- aspects to do with group composition, exploring, for example, relationships between group size or gender balance within groups and development of conceptual understanding
- the effectiveness of small-group discussions for different learning outcomes (e.g. attitude, decision-making)
- the nature of small-group discussions

Primary research

One particularly strong feature, which has emerged from the work undertaken for the review, is that there is a dearth of systematic research on small-group discussion work and considerable uncertainty on the part of teachers as to what they are required to do. Both these factors point to a pressing need for a medium- to large-scale research study which focuses on the use and effects of a limited number of carefully-structured, small-group discussion tasks aimed at developing various aspects of students' understanding of evidence.

1. BACKGROUND

1.1 Aims and rationale for current review

This review builds on work undertaken for an earlier systematic review by this RG (Bennett *et al.*, 2004) by continuing to focus on aspects of small-group discussion in science teaching. This area has been identified through consultation with groups, including science teachers, education researchers, teacher educators, curriculum developers and textbook writers, science inspectors, and professional organisations, all of whom are represented in the Review Group for Science. All members of the Review Group are in agreement that this area is extremely topical and of interest to a wide range of people involved in science education.

The overall review research question remains as it was for the initial review:

How are small-group discussions used in science teaching with students aged 11-18, and what are their effects on students' understanding in science or attitude to science?

This review led to a systematic map of the area and a first in-depth review of studies addressing the question: *What is the evidence from evaluative studies of the effect of small-group discussions on students' understanding of evidence in science?*

The systematic mapping of the area undertaken in the initial review revealed a wide range of relevant studies and facilitated the potential to explore a number of different aspects of the use of small-group discussion work in science teaching. One of these aspects was the ways in which different stimuli (printed materials, practical work, computers, etc.) are used to promote small-group discussion. As the literature in this area is extensive, particularly for the use of computers, it was decided to focus on how different stimuli are used to enhance students' understanding of evidence.

The review reported here therefore addresses the question:

What is the evidence from evaluative studies of the effect of using different stimuli (print materials, practical work, ICT, video/film) in small-group discussions on students' understanding of evidence in science?

1.2 Definitional and conceptual issues

The two most important definitional issues in the review concerned reaching an agreement on what constituted a *small-group discussion* and, for the in-depth review, what the term *evidence* would be taken to mean in science teaching.

Following discussion at a Review Group meeting, the following characteristics were agreed to be necessary for *small-group discussions*:

- They involve groups of two to six students.
- They have a specific stimulus (for example, a newspaper article, video clip, prepared curriculum materials).
- They involve a substantive discussion task of at least two minutes.
- They are either synchronous (that is, face-to-face) or asynchronous (that is, mainly IT-mediated).
- They have a specific purpose (for example, individual sense-making, leading to an oral presentation or to a written product).

Each of these aspects was incorporated into the review-specific keywords.

The term ‘evidence’ has become widely used in a number of educational contexts. In school science teaching, the notion of students’ use of evidence has its origins in the UK in the original version of the National Curriculum for Science, introduced in 1988, where one of the original 17 attainment targets focused on the history and development of ideas in science. Subsequent changes to the National Curriculum for Science saw the term ‘evidence’ being used in connection with investigative practical work, where students are required to support their results and conclusions with evidence based on the data they have collected. The most recent version of the National Curriculum (Department for Education and Skills (DfES), 1999) requires students to be taught about ideas and evidence in science. This move has served to focus attention on how students can be introduced to the notion of evidence in science lessons.

For the purposes of this review, the term ‘evidence’, in the context of school science teaching, has been taken to apply to activities which involve students in any of the following:

- engaging with data from primary and secondary sources (some of which may have been gathered by the students themselves)
- developing ideas in the form of claims or arguments
- drawing on the data to justify their claims or arguments

1.3 Policy and practice background

Interest in small-group discussion work in science

Small-group discussions have been strongly advocated as an important teaching approach in school science for a number of years. The use of small-group discussions in mainstream school science teaching has its origins in the widespread student-centred learning movement of the 1970s and 1980s, and in the development of context-based approaches to the teaching of science, where small-group discussion work was advocated as one of a range of teaching strategies seen as a means of helping students develop their scientific understanding.

Small-group discussion work and policy in science teaching

Although small-group discussion work is now strongly advocated for a number of reasons in school science teaching (see section 1.4), there has, until comparatively recently, been little formal policy on their use. However, concern in England and Wales over the suitability of the current science curriculum for the majority of 14- to 16-year-olds, has resulted in the development of a new science course for this age range, *21st Century Science* (www.21stcenturyscience.org). This course is aimed at developing students' scientific literacy, and small-group discussion work is seen as a key teaching strategy in this context. *21st Century Science* began its pilot phase in schools in September 2003; its outcomes will be central to shaping policy in future revisions of the school science curriculum. Thus it is likely that small-group discussion work will be advocated as policy in school science teaching, making a review of research in the area particularly timely.

1.4 Research background

Several factors have contributed to the current high levels of interest in small-group discussion work. These are summarised below. Some of the factors have emerged directly from research studies, while others appear to draw more loosely on research evidence and take the form of approaches which are being advocated in science teaching, but whose effects have yet to be explored on a more systematic basis.

The development of scientific literacy

The publication of *Beyond 2000* (Millar and Osborne, 1998) stimulated discussion and debate over the nature of the school science curriculum and, in particular, the ways in which it might foster the development of *scientific literacy*. This term embraces 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. There was also a clear message in the report of the House of Commons Science and Technology Committee (House of Commons, 2002) that scientific literacy would form part of a revised National Curriculum for Science: 'A new National Curriculum should require all students to be taught the skills of scientific literacy and selected key ideas across the sciences' (p 5).

A key aspect of scientific literacy is the ability to participate in informed discussion and debate of scientific issues, and this points to the need for including small-group discussions in the repertoire of activities employed in science lessons. Indeed, small-group discussions form a key teaching strategy in two new courses specifically aimed at developing scientific literacy: *Science for Public Understanding* (Hunt and Millar, 2000), a post-compulsory course for 17-and 18-year-olds, and *21st Century Science*, a GCSE course currently being developed by the University of York and the Nuffield Curriculum Centre.

Ideas about evidence

An area related to the development of scientific literacy is that of *ideas about evidence* (see also section 1.2): encouraging students to evaluate, interpret and analyse evidence from primary and secondary sources in science, including stories about how important science ideas were first developed and then established and finally accepted. This has led to considerations of the role of *argument* in school science, in the sense of putting forward claims and supporting them with sound and persuasive evidence (Osborne *et al.*, 2001). This has strong links with the use of small-group discussions, since the practice of using evidence in argumentation requires interaction with peers.

The constructivist viewpoint

One of the most significant research programmes in science education has emerged from the *constructivist viewpoint* on learning, which has explored in depth the ideas and understanding students bring with them to science lessons and the ways in which some of their ideas may hinder the development of accepted scientific ideas (e.g. Driver *et al.*, 1985). One of the recommendations for practice which has emerged from constructivist research is that small-group discussions should be used in science lessons as a means of helping students explore their ideas and move towards more scientific ideas and explanations. Further impetus for the inclusion of small-group discussions in science lessons has come from the development of ideas about *social constructivism* (Driver *et al.*, 1994). These draw on the work of Vygotsky who emphasises the importance of the social dynamics of interactions in fostering learning.

Formative assessment

The area of formative assessment is receiving considerable attention at present. Formative assessment relates to the assessment strategies and techniques which take place during teaching in order to establish progress and diagnose learning needs to support individual students. (This contrasts with summative assessment, which refers to the tests and examinations which take place at the end of courses or blocks of teaching.) A number of approaches have been advocated for increasing the use and effectiveness of formative assessment in science teaching, including the use of peer-review of work through small-group discussions (see, for example, Daws and Singh, 1999).

Learner-centred teaching and 'active learning'

Small-group discussions have been advocated for a number of years as one of a range of learner-centred teaching approaches or 'active learning' strategies. These terms are applied to activities in which students have a significant degree of autonomy over the learning activity, and are frequently advocated in teaching generally (for example, Kyriacou, 1998) and in science lessons specifically (for example, Bentley and Watts, 1989) as a means of stimulating students' interest in what they are studying.

Citizenship

In England and Wales, the notion of citizenship currently has a very high profile. In October 2002, it became a compulsory component of the National Curriculum, to be addressed within other school subjects. While discussion and debate over what comprises citizenship are still ongoing, it is clear that there are links with scientific literacy, as the latter seeks to provide young people with the information and skills they need to help them think and act appropriately on scientific matters which may affect their lives as future adult citizens. Thus small-group discussions have a role to play in the context of citizenship as part of the school curriculum.

The development of literacy skills

There is a more general drive to improve students' literacy skills and, in England and Wales, this has been formalised into the National Literacy Strategy (DfEE, 1998). Small-group discussions have been advocated as a means for developing students' language skills in science (e.g. Newton *et al.*, 1999, and Osborne *et al.*, 2001).

Research into the use of small-group discussion work

There is a growing body of evidence that teachers would welcome support and guidance on running small-group discussions (for example, Newton *et al.*, 1999).

In particular, evaluation work undertaken on materials and courses with a specific focus on teaching socio-scientific issues and developing scientific literacy, the new *AS Public Understanding of Science* course (Osborne *et al.*, 2002) and the *Valuable Lessons* project (Levinson and Turner, 2001), established that teachers saw the provision of support and guidance on running small-group discussions as a priority. In addition, the evidence from another systematic review of small-group discussion work (Bennett *et al.*, forthcoming) adds further strength to the need for support for teachers. While the ability to engage in discussion is seen as an important part of the science education of young people, science-based learning activities aimed at developing this ability are not well known to science teachers. Furthermore, the introduction of small-group discussions in science lessons challenges the established pedagogy of science teaching and places new demands on science teachers.

Taken together, the factors outlined above pointed very strongly to the desirability of the first review of the use of small-group discussions in science teaching (Bennett *et al.*, 2004). There are two reasons for choosing to continue with work in this area. First, the searching undertaken for the first review yielded a systematic map of some 90 studies, making it impractical to explore all these in the in-depth review. Second, the focus remains very topical and the Review Group has had a number of approaches from different groups (e.g. the project team working on the new GCSE science course, *21st Century Science*) interested in the findings of the review.

A note on collaborative learning

There is a large quantity of mainly USA-based literature on collaborative learning, which at first sight would appear to be of direct relevance to small-group discussion work, in that one would assume that discussion formed part of the majority of tasks set in a collaborative learning situation. Certainly this term was included in the electronic searches. However, closer examination of the literature indicated that the focus was primarily on *strategies* to promote collaborative learning. Little, if any, *direct* reference was made to small-group discussion work, although, by implication, it must have been taking place. It was therefore decided that, for the purposes of the research review question, this area of work would be excluded unless reference was made to the use of specific discussion tasks and their effects.

A number of collaborative learning strategies are described briefly below, as they clearly involve students discussing ideas, and are therefore useful starting points for the development of materials aimed at promoting small-group discussion work.

Jigsawing: Jigsawing involves students in being members of two different groups (Aronson *et al.*, 1978). The first is the ‘home’ group, in which students work in groups of four to six on some instructional material which has been broken down into sections. Each student in the home groups is assigned a different portion of the material. The home groups then break apart and re-form into ‘expert’ groups in which group members all focus on, and discuss, the same piece of the material to make sure they understand it. Once this has happened, students’ groups then break once again and re-form back into ‘home groups’ to peer-tutor the home group on the aspect of the material they have studied intensively, and learn from other home group members about the other aspects of the material.

Envoying: This technique also involves students working in two groups. In the first group, they discuss a common task, which differs for each group. Groups then re-form, with new groups containing one member of each of the original groups, who act as envoys to report on their particular task.

Snowballing: In a ‘snowball’ exercise, pairs of students discuss a question or idea and agree on their views, then join with another pair to share what they have discussed, and then finally with another group of four (two pairs) to share thinking for a final time.

Four corners: The teacher chooses a topic and the students then brainstorm related sub-topics. Through a process of elimination, four topics are identified and one each is allocated to students grouped into the four corners of the room. The groups then choose a leader, a recorder and a reporter. The topics are discussed in the groups and the reporter then summarises them for the other groups.

1.5 Authors, funders and other users of the review

The review is being undertaken by this Review Group because its members have both expertise and interest in the area of small-group discussion work, as well as experience of undertaking systematic review work. As described above, the review focus – small-group discussion work in science – is particularly topical at present, being of central concern to policy-makers, teachers, advisory teachers, inspectors, academic researchers, teacher trainers and those involved in curriculum development work. The Review Group membership reflects the various constituencies interested in small-group discussion work in science education.

2. METHODS USED IN THE REVIEW

2.1 User involvement

2.1.1 Approach and rationale

The Review Group contains representatives from most of the key constituency groups in science education (lead teachers, teacher educators, curriculum developers, educational advisers and inspectors, policy-makers and academics) in the area of science and science education.

2.1.2 Methods used

All group members have been involved in most key stages of the review, including:

- the decision over the review question(s)
- the development of inclusion and exclusion criteria
- the development of review-specific keywords
- the identification of the focus for the in-depth review

The accelerated deadline for the completion of this in-depth review report precluded all group members commenting on the final content of this report.

School students are also a key constituency group. While it is impractical to invite them to attend Review Group meetings, they will be involved in commenting on the findings of the review. All teacher members of the Review Group indicated the feasibility of involving their students in the review and a willingness to assist with this aspect of the review.

A further group of review users are teachers in training. Funding was secured to involve Postgraduate Certificate in Education (PGCE) students in producing user-friendly summaries of the first review findings for teachers, teacher educators and students. This formed part of their regular training programme. The product was distributed amongst key-members of the respective target groups through the University of York Science Education Group (UYSEG) network and the Association for Science Education (ASE).

The Review Group also benefited from the advice of a group of national and international consultants, all with expertise in particular aspects of science education, and including the editors of the major international science education journals. One purpose of establishing such a group was to ensure that the review had an international perspective. Members of this group have been consulted over the suitability of the research review question and acted as key informants in providing the Review Group with details of any work they saw as suitable for potential inclusion in the review.

Appendix 1.1 lists the members of the Consultancy Group.

2.2 Identifying and describing studies

A research study may be reported in a number of research 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. The map of research is presented as an overview of characteristics of research studies, where applicable, based on keywords of a combination of papers reporting the same study.

2.2.1 Defining relevant studies: inclusion and exclusion criteria

For the second review focusing on aspects of small-group discussion work in science teaching, the same inclusion/exclusion criteria were used as in the first review. An important exception was that the period covered was extended from 1980–2002 to include 2003. This allowed the map to be updated.

The EPPI-Centre systematic review methods were followed for searching, screening and including (or excluding) studies in the map, and in applying the EPPI-Centre keywording sheet and keywording strategy (EPPI-Centre, 2002a, 2002b), supplemented by review-specific keywords, to studies. Extracting data and making quality assessment of studies included in the in-depth review were also carried out according to EPPI-Centre procedures and using EPPI-Centre software (EPPI-Centre, 2002c).

The systematic map was based on studies identified in the first review. Additional studies were included if they met the following criteria:

- They were about the use of small-group discussions in science lessons.

- They involved groups of two to six students.

- They involved a substantive, structured discussion task of more than one or two minutes duration.

- They illustrated how small-group discussions are being used.

- They focused on learning outcomes, addressed aspects of students' understanding in science or addressed aspects of students' attitudes to science.

- They were empirical studies of the following types: descriptive, exploration of relationships, evaluation (naturally-occurring and researcher-manipulated), or reviews (systematic and non-systematic).

- They were about students in the 11–18 age range.
- They had been undertaken in the period 2003.
- They were published in English.

For a justification of these inclusion criteria, the reader is referred to Bennett *et al*. (2004).

Studies are *excluded* from the systematic map if they are not about science, not about relevant aspects of science, not of specified study type, not within the specified age range and not within the specified period.

The inclusion and exclusion criteria are set out in detail in Appendix 2.1.

2.2.2 Identification of potential studies: search strategy

The search strategy for this in-depth study was to use the relevant studies already identified in the first in-depth strategy and to update them to include papers published in 2003. The same methods of electronic and handsearching, and personal approaches were employed with two modifications. No search of PsycINFO was included as experience in the previous review indicated that no papers, other than those already found through ERIC and SSCI, were identified. Second, by 1st May 2004, the ERIC database had only been partially updated for 2003. The searching string for BEI was, therefore, enlarged compared with the search string for the initial review. Appendix 2.2 gives details of the search strategy terms used for the electronic databases.

The respective inclusion criteria define studies to be included in the review as:

1a They are about group discussions
1b which take place in science lessons.
2 The groups should have two to six participants.
3a Group discussions should be based on a structured task
3b and take more than two minutes.
4a They address aspects of students' understanding of science
4b or aspects of students' attitudes to science.
5a They are empirical descriptive explorations of relationships
5b or they are empirical evaluations
5c or they are reviews.
6a They are about students
6b in the age range of 11–18 years.
7 They have been published in the period 1980–2003.
8 They are published in English.

2.2.3 Screening studies: applying inclusion and exclusion criteria

The Review Group set up a database system (using EndNote software) for keeping track of and coding papers found during the review. Titles and abstracts were imported electronically and entered manually into the review database as appropriate. Inclusion and exclusion criteria were applied successively to (i) titles and abstracts, and (ii) full reports. Papers excluded on the basis of titles and abstracts were recorded on the database with reasons for their exclusion. Excluded papers of potential interest for theoretical and policy background were marked as such. Full reports of potentially relevant studies were obtained from the University of York library or sent for through interlibrary lending. Inclusion and exclusion criteria were re-applied to the full reports and those which did not meet these initial criteria were excluded. At both stages of screening, the inclusion and exclusion criteria were applied hierarchically, such that, for instance, exclusion on criterion 6 implied that the study met the inclusion criteria 1–5. The database was fully annotated with reviewer decisions on inclusion and exclusion and reasons for exclusion.

2.2.4 Characterising included studies

The studies remaining after application of the criteria were keyworded, using the EPPI-Centre generic keywording sheet and keywording strategy (EPPI-Centre, 2002a, 2002b). Additional keywords specific to the context of the review were added to those of the EPPI-Centre. (Appendix 2.4 gives details of the generic and the review-specific keywords.)

A systematic map of the research in the field was drawn using the generic and review-specific keywording sheets. This is presented in Chapter 3 in the form of narrative and mapping tables scrutinising the following areas:

- publication date of studies
- country of origin
- study type
- science discipline
- types of learners
- number of students
- constitution of discussion groups
- duration of discussion tasks
- stimulus for discussion tasks
- product of discussion tasks
- outcomes reported
- number of discussion groups
- research strategy used
- nature of data collected
- relationships between discussion stimulus and reported learning outcomes

2.2.5 Identifying and describing studies: quality-assurance process

EPPI-Centre procedures require that a percentage of papers are double-screened for quality assurance purposes. This approach was adopted for the 1980 – 2002 papers (see previous report, Bennett *et al*., 2004) when papers were first screened on title and abstract. To check for consistency of coding 2.5% of papers (45 chosen randomly) were initially screen independently by all four members of the Science team and a further 2.5%, also chosen randomly, screened independently by all the team members and a member of the EPPI-Centre. Once consistency was established, the rest of the papers were divided amongst three members of the team for screening. A further round of checks was applied to a second round of screening when full papers were available for detailed inspection.

In this review, all 249 papers published in 2003 and found by the various search routes were independently screened by two people, their decisions then discussed and agreed. The same team members screened 18 papers during the second screening cycle.

These data were used to calculate inter-screener agreement, using frequency counts and the Cohen's Kappa inter-screener reliability coefficients. Details of these are given in section 3.3.

The same keywording process was carried out as for the previous review.

2.3 In-depth review

2.3.1 Moving from broad characterisation (mapping) to the 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 included studies.

In the light of what emerged in the systematic map, and on the advice of the Review Group, the review question was refined for the in-depth review as:

What is the evidence from evaluative studies of the effect of different stimuli (print materials, practical work, ICT, video/film) in small-group discussions on students' understanding of evidence in science?

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

1. Exclusion on study type (that is, the study is not an evaluation, either naturally-occurring or researcher-manipulated)
2. Exclusion on study focus (that is, the study is not an evaluation of the effect of stimuli)
3. Exclusion on study outcome (that is, the study does not report on change in students' understanding of evidence in science)

For the purposes of this review, 'understanding of evidence' was defined as the understanding 'related to the collection, validation, representation and interpretation of evidence' (Gott and Duggan, 1996, p 793); that is, the ability to co-ordinate observations (primary or secondary data) with theory (models or concepts). We excluded studies that focused on outcomes such as 'conceptual understanding of science concepts', 'applications of science', 'attitudes to (school) science', 'communication or collaboration skills', or 'decision-making skills on socio-scientific issues', as identified through the review-specific keywording sheet in Appendix 2.4.

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 software, EPPI-Reviewer (EPPI-Centre, 2002c).

2.3.3 Assessing quality of studies and the 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 in-depth review question. The EPPI-Centre data-extraction procedures identify three weightings – high, medium and low – to help in the process of apportioning different weights to the findings of different studies. For the purposes of this review, we have refined these weightings as follows: high, medium-high, medium, medium-low and low.

Weightings are given to the following categories:

Category A: The trustworthiness of findings (internal methodological coherence) in relation to the study's own research question(s)

Category B: The appropriateness of the research design and analysis used for answering the in-depth review question

Category C: The relevance of the study topic focus (from the sample, measures, scenario, or other indicator of the focus of the study) to the in-depth 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, a judgement of quality within the EPPI-Centre data-extraction guidelines (EPPI-Centre, 2002d) was used (M.11).

Judgements of weighting in categories B and C are based on the quality of the study's research work solely related to the in-depth review question. Appendix 2.5 shows how the Review Group interpreted the appropriateness of the research design and analysis (category B) through five aspects: the sample size/sampling method; nature of a comparison group; benchmark data; 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 weighting 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 2.5 shows how the relevance of the study topic focus (category C) has been weighted through five aspects: nature of small-group discussion (how representative); the relative importance of the stimulus in the intervention; the appropriateness of the measures for testing understanding of evidence; the breadth of understanding of evidence measured; and the representativeness of the study situation (learners in the classroom). Again, each of these aspects has three level descriptors with weighting 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.

2.3.4 Synthesis of evidence

The final step in the review is to synthesise the findings and bring together the studies which answer the review questions and which meet the quality criteria relating to appropriateness and methodology.

For each study, a summary report (see Appendix 4.1) was drawn up, using key items within the EPPI-Reviewer data extraction tool. These items were agreed amongst the core Review Group. Only one characteristic considered important was not included in this tool – the 'details of the researchers' – and this information is included in the summary tables. These reports were edited by one group member for consistency of terminology, depth and detail, continuously referring to each relevant study. The reports were used by two group members to identify commonalities across the studies for the same characteristics as presented in the map. In addition, commonalities of, and differences between, studies were identified for methodological aspects of the studies on the basis of these reports. The latter resulted in the judgement of 'weight of evidence A'. For the synthesis of the appropriateness of the studies' research design and analysis (weight of evidence B), the five characteristics listed in weight of evidence B were used as organisers. The same was the case for the synthesis of the relevance of the focus of the studies (weight of evidence C). This synthesis method necessitated a continuous consultation between two group members. There was a strong interplay between the synthesis of methodological characteristics, and judgements made on the basis of these characteristics, thus improving the consistency of the weightings for the set of studies.

The consolidated evidence from this review draws on the findings from studies weighted as *high*, *medium-high* and as *medium*, as summarised above.

Short summaries of the relevant results from each of the studies that were judged to be of sufficient quality are given in Table 4.7. The findings from these individual studies were then clustered according to common features agreed by two members of the group. The emerging themes are described and discussed in section 4.4 as the findings of this review.

2.3.5 In-depth review: quality-assurance process

Data extraction and assessment of the weight of evidence brought by the study to address the review question were conducted by pairs of Review Group members working first independently and then comparing their decisions and coming to a consensus. In addition, for purposes of quality assurance, a member of the EPPI-Centre double data-extracted and quality assessed one of the papers included in the in-depth review.

3. IDENTIFYING AND DESCRIBING STUDIES: RESULTS

3.1 Studies included from searching and screening

Figure 3.1 provides a summary of the number of papers and studies involved at various stages of the filtering process. The process of searching yielded 2,246 papers, 249 of which were identified by updating the period covered to include 2003. An additional 44 papers were identified through handsearching or personal contacts; thus the review handled a total of 2,290 records. After de-duplication and the first round of screening 391 papers remained as potentially included. Hard copies of only 12 papers (3%) were unobtainable. After second screening, 119 papers remained for inclusion in the review. Papers reporting on the same study were identified as described at the beginning of section 2.2. The 119 papers were found to report on 94 studies, 10 of which were included in the in-depth review.

3.2 Characteristics of the included studies (systematic map)

Table 3.1: Publication date of studies included in the systematic map (N = 94, mutually exclusive)

Publication period	Number of studies	%
1980 – 1985	1	1
1986 – 1991	5	6
1992 – 1997	38	40
1998 – 2003	50	53
Total	***94***	***100***

Table 3.1 indicates that the research activity in the review area has been minimal up to ten years ago and has most prolific in the last five years. It also demonstrates that the research area under review is currently still very active, and likely to be relevant to a considerable number of researchers, research policy-makers and others.

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

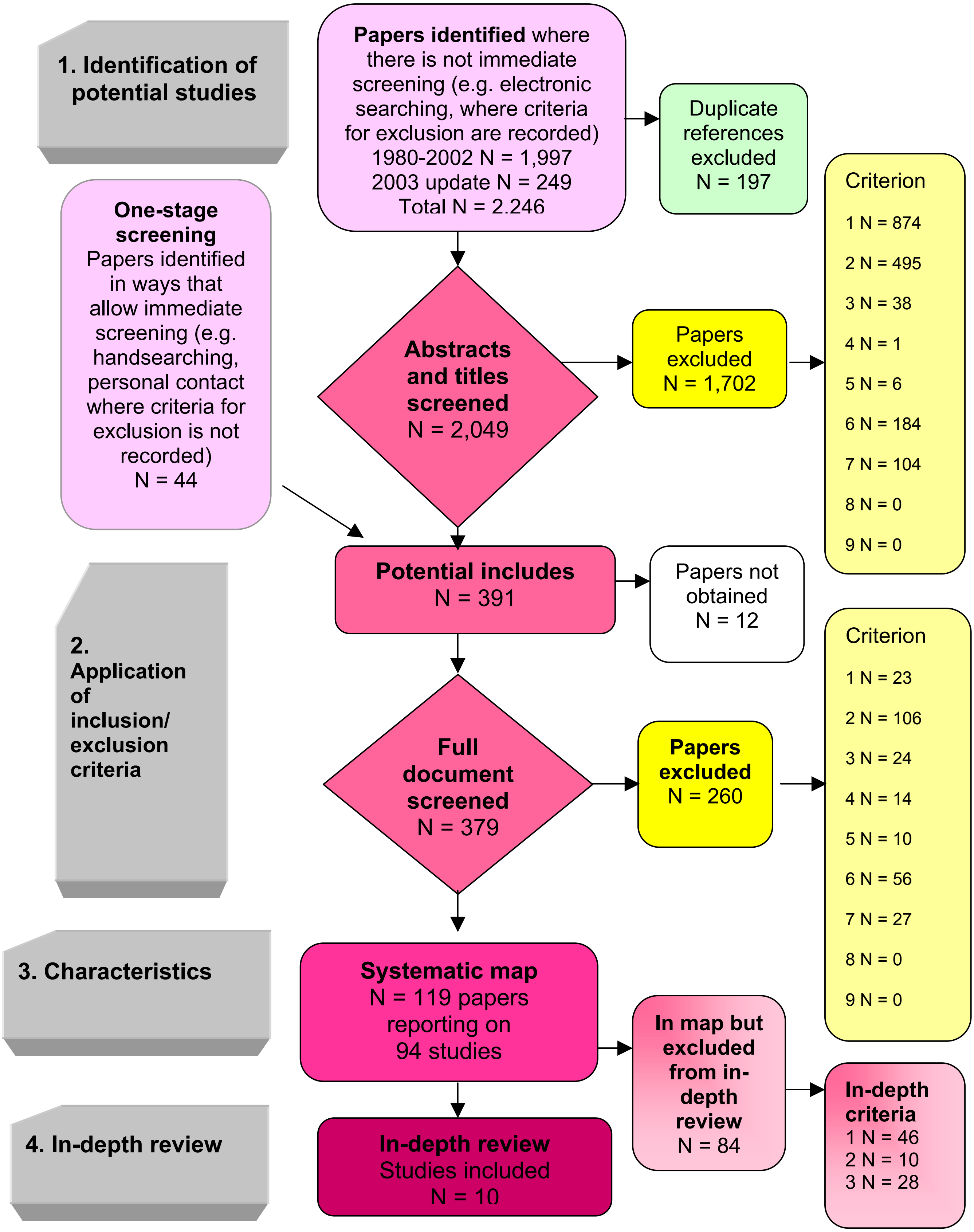
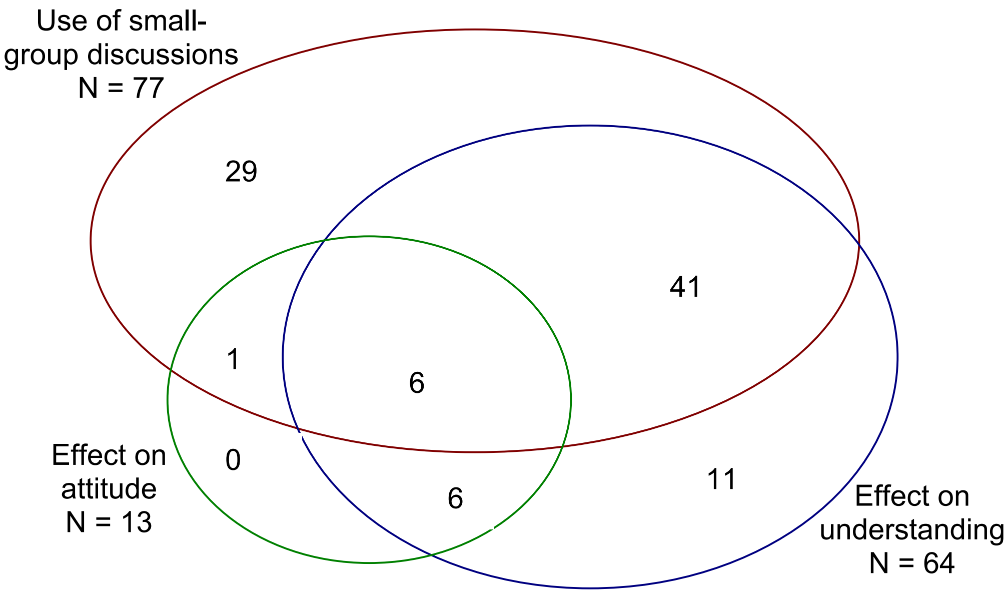

1. Identification of potential studies
Papers identified where there is not immediate screening (e.g. electronic searching, where criteria for exclusion are recorded)
1980-2002 N = 1,997
2003 update N = 249
Total N = 2.246
Duplicate references excluded N = 197
One-stage screening
Papers identified in ways that allow immediate screening (e.g. handsearching, personal contact where criteria for exclusion is not recorded)
N = 44
Abstracts and titles screened N = 2,049
Papers excluded N = 1,702
Criterion
1 N = 874
2 N = 495
3 N = 38
4 N = 1
5 N = 6
6 N = 184
7 N = 104
8 N = 0
9 N = 0
Potential includes N = 391
Papers not obtained N = 12
2. Application of inclusion/ exclusion criteria
Full document screened N = 379
Papers excluded N = 260
Criterion
1 N = 23
2 N = 106
3 N = 24
4 N = 14
5 N = 10
6 N = 56
7 N = 27
8 N = 0
9 N = 0
3. Characteristics
Systematic map N = 119 papers reporting on 94 studies
In map but excluded from in-depth review N = 84
In-depth criteria
1 N = 46
2 N = 10
3 N = 28
4. In-depth review
In-depth review Studies included N = 10

Figure 3.2: Focus of the included studies (N = 94)



NB: Venn diagram is not to scale.

The review question has three components. The first component focuses on the process of what takes place during small-group discussions, in short the *use of small-group discussions*. The remaining two components focus on outcomes of small-group discussions: that is, the effect on group members' *understanding of science* and on their *attitude to (school) science*. Figure 3.2 indicates the focus of the 94 studies included in the review. Not surprisingly, the majority of studies (77) report on the process of small-group discussions, although only 29 of these solely report on this aspect. Just over half of these studies (41) also report on the effect on students' understanding of science. A total of 64 studies report on students' understanding of science, with only eleven of these solely dealing with this aspect. A small number of studies (13) report on the effect of small-group-discussions on students' attitude to science, with half of these (six) reporting on all three aspects of the review question.

Table 3.2: Country in which the study carried out (N = 94, not mutually exclusive)

Country	Number of studies	% of the 94 studies
USA	37	39
UK	12	13
Canada	11	12
Australia	6	6
Germany	5	5
Hong Kong	5	5
Netherlands	4	4
Taiwan	4	4
Spain	3	3
Finland	2	2

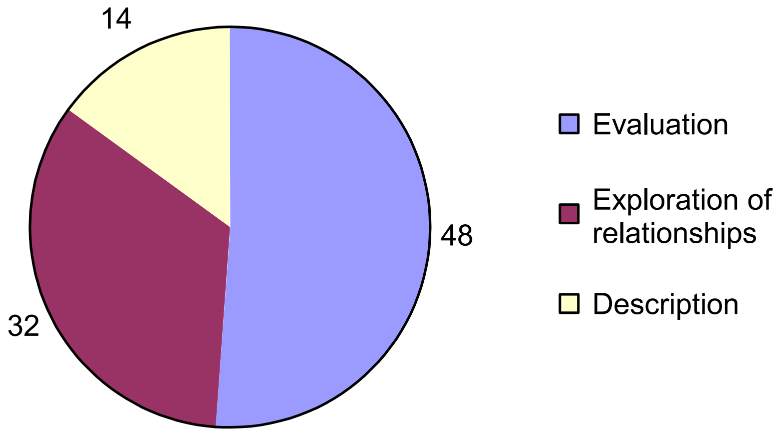
Country	Number of studies	% of the 94 studies
France	2	2
Israel	2	2
Greece	1	1
Malaysia	1	1
Brazil	1	1
Singapore	1	1
Total	***97***	

Data in Table 3.2 demonstrate that the 94 studies included in the map report on studies carried out in a large number of different countries, even though the review was focused on papers published in English. Two studies draw on data from three and two countries, respectively. As this review was limited to publications in English, one would expect that studies in English-speaking countries might be over-represented. Compared with other systematic reviews – for instance, a review of research on the effects of context-based approaches to learning (Bennett *et al.*, 2003) – a proportion of about two-thirds of studies from the US, UK, Canada and Australia is not unusual. In addition, the review does include studies of small-group discussions held in Bahasa Malay, Cantonese, Dutch, Finnish, French, German, Greek, Hebrew, Mandarin, Portuguese and Spanish. It is of note that no studies focus on small-group discussions of learners talking in English as their second language or who are hearing and/or speech impaired.

The large number of studies from the USA reflects various active research groups. Using the inclusion of at least three studies in this review as a yardstick, very productive research in the review area takes place at the Universities of Michigan (by Anderson, Palincsar, Vellom), Miami (by Roychoudhury) and at the Institute of Ecosystem Studies, NY (by Hogan). In contrast, the studies from the UK seem to stem mainly from individuals and not research groups with only Osborne (Kings College London) publishing at least three of the studies in the review. The Canadian studies reflect very productive work at the University of Victoria (by Roth). Equally, there are specialist researchers in the review area in Hong Kong (Tao) and Spain (Jimenez-Aleixandre).

All studies have a topic focus at the interface of the curriculum and teaching/learning strategies in the curriculum area of science. The majority of studies (84) focus on secondary school learners between 11 and 16 years of age since studies with younger age groups were excluded from the review. Eighteen also report on the age range 17 to 20. The learners in most samples (85) were of mixed sex. A total of four and 11 studies report on female- and male-only educational settings respectively.

Figure 3.3: Study type (N = 94)



The EPPI-Centre uses a system of classifying types of research studies. A study may solely provide a description of a process. It may, in addition, identify relationships between different characteristics of a process. Finally, it may focus on an intervention and evaluate this against specific outcomes. Many reports of evaluative studies also explore relationships and provide descriptions of processes. For our review, Figure 3.3 indicates that more than half (48) of the studies report on evaluation studies, split almost equally between naturally-occurring (23) and researcher-manipulated (25) evaluations of the effect of small-group discussions. Of the latter, 12 studies report a RCT. Just over one-third of the studies (32) present explorations of relationships between different characteristics of small-group discussions. A minority of studies (14) provide only descriptions of small-group discussions.

Table 3.3: Distribution by discipline (N = 94, not mutually exclusive)

Science subject	Number of studies	% of the 94 studies
(Integrated) science	37	39
Biology	18	19
Chemistry	4	4
Physics	37	39
Earth science	4	4
Total	***100***	

Table 3.3 shows that the review involves a large number of studies of small-group discussions in science and physics, a smaller number in biology and very few in chemistry and earth science. Most of the small-group discussions developing skills of decision-making on socio-scientific issues are, traditionally, placed within biology classes. The difference in the frequency of studies in physics and chemistry classes is surprising, as the nature of small-group discussions in both subjects may not be very different. This difference could be explained by the fact that the subject background of many productive researchers in this area is physics, rather than chemistry. Alternatively, small-group discussions may indeed be more prominent in physics, whereas any perceived need for learner-led classroom activities in chemistry may be satisfied in other ways, such as through the use of practical work.

Table 3.4: What types of learners are involved? (N = 94, not mutually exclusive)

Ability of learners	Number of studies	% of the 94 studies
Mixed ability	79	84
Lower ability/slow learners	4	4
Upper ability/gifted	13	14
Disaffected	2	2
Unspecified	2	2
Total	***100***	

Several authors did not specify the ability level of the learners reported. However, familiarity with the comprehensive nature of most of the education systems involved allowed reviewers to infer ability levels in all but two cases.

Table 3.4 indicates that the majority of studies involved mixed ability classes, sometimes grouped in homogeneous ability discussion groups. A small cluster of studies report on small-discussion groups in high ability learners.

Table 3.5: Number of students in a discussion group (N = 94, not mutually exclusive)

Group size	Number of studies	% of the 94 studies
Groups of 2 (dyads)	31	33
Groups of 3–4	58	62
Groups of 5–6	14	15
Unspecified	10	10
Total	***113***	

Table 3.5 indicates that one-third of the studies focus on groups of two students, and almost double that number focus on groups of three or four learners. It is noted that, especially for samples of lower ability or disaffected students, a statement of group size was often omitted as the group size was usually unstable.

Table 3.6: How discussion groups are constituted (N = 94, not mutually exclusive)

Group composition method	Number of studies	% of the 94 studies
Friendship ties	15	16
Randomly, by teacher	11	12
Randomly, but same sex groups	6	6
Purposely same ability	5	5
Purposely heterogeneously	30	32
Unspecified	40	43
Total	***107***	

The way discussion groups were constituted is given in Table 3.6.

Practitioners of co-operative learning strategies are often very specific about the composition of discussion groups. This might relate to students being of similar or different abilities or ways of thinking: for example, in the study by De Vries *et al.* (2002, p 97), students were tested for the conceptual models they used and then paired in ways to make them more likely to engage in discourse. In other studies, students were allowed to form friendship groups as this is considered to encourage discussion (Gayford, 1995, p 136).

About 16% of all studies (15) allow groups to emerge from friendship ties and 43% of all studies (40) do not specify the way groups are constituted. Several of these are likely to allow friendship ties as a base for group composition. As a consequence, it was estimated that only one-third of the groups were deliberately constituted.

How groups are organised

Co-operative learning strategies also differentiate between various ways discussion tasks are organised. These include snowballing, where students started work in pairs, then worked in groups of four (for example, Taconis and Van Hout-Wolters, 1999, p 317; Pedersen, 1992, p 375) and jigsawing, where small groups of cooperating students treat each other as a resource and change groups to exchange knowledge gained in their first group (for example, Lazarowitz *et al.*, 1988, p 477). However, almost all the studies in this review (89) concern studies of self-contained and permanent groups, with only three studies considering snowballing and two jigsawing as an organisational structure.

Table 3.7: Duration of the discussion tasks (N = 94, not mutually exclusive)

Duration of the discussion tasks	Number of studies	% of the 94 studies
2–5 minutes	1	1
6–30 minutes	9	10
Close to class period (30–60 minutes)	32	34
Longer than a class period	32	34
Unspecified	24	27
Total	***98***	

The duration of the discussion tasks reported in the studies often had to be inferred from the reported length of the tape-recording or activity. The description of the research method in just over one-quarter of the studies (24) does not allow such inferences. Somewhat surprisingly, Table 3.7 shows that two-thirds of the studies (64) report group discussions lasting close to, or exceeding, a class period. This seems a long, probably unrealistic, period for meaningful discussion, unless it takes place in the context of a project, practical activity or poster construction.

Table 3.8: Stimulus for discussion tasks (N = 94, not mutually exclusive)

Nature of the stimulus for the discussion task	Number of studies	% of the 94 studies
One-line oral teacher instruction	2	2
Oral context provided by teacher only	3	3
Newspaper article	1	1
Prepared curriculum print materials	62	66
Practical work	37	39
Computer software	25	27
Field trip	1	1
Video/TV/film clip	8	9
Learner generated	14	15
Total	***153***	

Discussion tasks in half of the studies used more than one type of stimulus (Table 3.8). In two-thirds of the studies (62), discussions were based on curriculum print materials, usually a worksheet, a handout with text, specific problems to be solved or issues to be discussed. In more than one-third of the studies (37), the group discussions centred around practical work, and in just under a quarter of the studies (25), the stimulus was computer software, just for display or interactive versions. Video, TV or film clips were used (mainly in the older studies) in fewer than one in ten of the studies in the review. It is notable that field trips and newspaper articles have hardly been reported as stimuli for group discussions.

The Review Group had a special interest in asynchronous discussions, typically using ICT at a distance. Although the search yielded a sizeable number of studies dealing with educational software facilitating asynchronous interaction of different students (for instance, in chatrooms or through designated project websites), only four studies were eventually included in the review. Many of the other studies did not focus on the group discussions resulting from the use of the software, but instead described the software features and the frequency of their use and accessibility in practice.

Table 3.9: Product of discussion tasks (N = 94, not mutually exclusive)

Product of the discussion task	Number of studies	% of the 94 studies
Individual sense-making	88	94
Report group views/presentation orally in class	21	22
Support a group position in a class debate/quiz	10	11
Present group written project (including poster)	11	12
Other	6	6
Total	***136***	

In a very high proportion of the studies, the product of the discussion task was individual understanding of the science underlying the activity, such as a practical experiment, the preparation of a poster or a computer-based exercise, in which they were engaged (Table 3.9). In just under half of the cases (42), this understanding was then shared with classmates in different ways: groups might present their findings or views orally or by way of posters, or might defend their position in a whole class debate. Those products falling into the 'other' category included either group or individual written reports or posters that were submitted to the teacher or researcher.

Table 3.10: What outcomes are reported (N = 94, not mutually exclusive)

Reported outcome	Number of studies	% of the 94 studies
Conceptual understanding of science	69	73
Evidence (methods and nature of science)	31	33
Applications of science	3	3
Attitudes to (school) science	15	16
Skills (communication/collaboration)	57	61
Decision-making on socio-scientific issues	10	11
Total	***185***	

As can be seen in Table 3.10, the reported outcomes in the studies often included more than one aspect per study. Nearly three-quarters (73%) of studies focused on the impact of the discussion tasks on the conceptual aspects of science understanding, while one-third (33%) were interested in the understanding of evidence. Not surprisingly over a half of the studies (61%) focused on the actual communication and collaborative skills associated with the discussion tasks involved in group work. A small proportion of studies involved decision-making on socio-scientific issues and very few included aspects relating to the applications of science.

Table 3.11: Number of discussion groups included in the study (N = 94, not mutually exclusive)

Number of groups in the study	Number of studies	% of the 94 studies
1 discussion group only	10	10
2 discussion groups	5	5
3–10 discussion groups	38	40
11–30 discussion groups	26	28
More than 30 discussion groups	15	16
Unspecified	5	5
Total	***99***	

The majority of studies involved three or more discussion groups with the highest number (38) being in the range 3-10 (Table 3.11). These studies would normally focus on a single class or a subset of groups within it. The distribution would be close to normal, except for the relatively large number (10) of those involving only one group. These studies usually report very detailed analysis of the discourse of a small-group of students carrying out a task. This approach is favoured in the various studies by Roth and colleagues.

Table 3.12: Research strategy used (N = 94, not mutually exclusive)

Research strategy	Number of studies	% of the 94 studies
Experiment	28	30
Survey	15	16
Case study	53	56
Action research	3	3
Ethnography	4	4
Total	***103***	

It is surprising that more than half of the studies (53) in the review can be characterised as case studies (Table 3.12). Just under one-third of the studies (28) use an experimental design: that is, a study with an experimental and a control group. This would include all researcher-manipulated evaluations and a selection of naturally-occurring evaluations. One in six (15 studies) constitute a survey.

Table 3.13: Nature of the data collected (N = 94, not mutually exclusive)

Nature of the data collected	Number of studies	% of the 94 studies
Test results	40	43
External examination results	1	1
Written reports/questionnaires	32	34
Concept webs	5	5
(Dis)agreement scores (e.g. VOSTS)	3	3
Self-reports (diaries, interviews)	34	36
Group discussions (audiotaped)	42	45
Presentations	1	1
Observed behaviour (including videotaped)	64	68
Computer logs	15	16
Total	***237***	

On average, the studies present findings based on more than two different types of data (Table 3.13). Half of the research reports on small-group discussions are based on audiotaped (42 studies) and/or videotaped (64 studies) interactions. In addition, almost half of the studies (40), especially those on evaluation studies, present data through attainment test results of discussion group members. Approximately one-third of the studies used questionnaires (32) and interviews (34) for collecting data.

Table 3.14: Relationships between discussion stimulus and reported learning outcomes (N = 94, neither category mutually exclusive)

Nature of the stimulus for the discussion task	**Total number of studies reporting each stimulus**	**Reported learning outcome**					
		Concepts	**Evidence**	**Application**	**Attitude**	**Communi-cation skills**	**Decision-making skills**
One-line oral teacher instruction	2	1	1	1	1	1	1
Oral context provided by teacher	3	3	-	-	-	-	-
Newspaper article	1	1	1	1	1	-	1
Prepared curriculum print materials	62	48	20	2	8	38	8
Practical work	37	22	12	-	6	25	-
Computer software	25	18	9	-	3	16	-
Field trip	1	1	1	-	1	1	-
Video/TV/film clip	8	7	3	-	4	6	1
Learner generated	14	11	6	-	2	9	-
Total	***153***	***112***	***53***	***4***	***26***	***96***	***11***

The cross-tabulation in Table 3.14 indicates that the small-group discussions in the various studies reported in the review studies show no particular focus on the type of stimulus in relation to the learning outcome that the study reports. In other words, the different types of stimulus are equally represented across the different learning outcomes researched. A cross-tabulation for the type of stimulus used in small-group discussions at different age levels equally does not indicate any preference for any specific type of stimulus.

Appendix 3.1 tabulates all 94 studies in the review according to the type of research study reported.

3.3 Identifying and describing studies: quality-assurance results

The quality-assurance processes for searching, screening and keywording described in section 2.2.5 were used with the following results.

The inter-screener reliability as measured by the frequency method and the Cohen’s Kappa method is shown in the Table 3.15. The Cohen’s Kappa method has the advantage of compensating for chance agreement.

Table 3.15: Inter-screener agreement (include-exclude) for first and second screening

	Frequency method		**Cohen's method**	
	Identical decisions	Inter-screener agreement	Cohen's Kappa coefficient	Inter-screener agreement
1st screening (N=249): Screener 1-Screener 2	246	98.8%	0.865	Very good
2nd screening (N=18): Screener 1 - Screener 2	17	94.4%	0.879	Very good

Percentage inter-screener agreement is at a very high level (98.8% and 94.4% for first and second screening respectively), and so is the Cohen's Kappa value (0.865 and 0.879). Any discrepancies between decisions of screeners 1 and 2 were discussed and resolved.

As a result of the screening process, five (new) studies were included in the review. These studies were keyworded independently by two team members, with an inter-coder agreement of 92.2%.

4. IN-DEPTH REVIEW: RESULTS

4.1 Selecting the studies for the in-depth review

The application of the exclusion criteria specified in section 2.3.1 resulted in ten studies for the in-depth review:

Keys CW (1997) An investigation of the relationship between scientific reasoning, conceptual knowledge and model formation in a naturalistic setting. *International Journal of Science Education* **19:** 957-970.

Lajoie SP, Lavigne NC, Guerrera C, Munsie SD (2001) Constructing knowledge in the context of BioWorld. *Instructional Science* **29:** 155-186.

Lavoie DR (1999) Effects of emphasising hypothetico-predictive reasoning within the science learning cycle on high school student's process skills and conceptual understandings in biology. *Journal of Research in Science Teaching* **36:** 1127-1147.

Palincsar AS, Anderson C, David YM (1993) Pursuing scientific literacy in the middle grades through collaborative problem solving. *Elementary School Journal* **93:** 643-658.

Sherman GP, Klein JD (1995a) The effects of cued interaction and ability grouping during co-operative computer-based science instruction. *Educational Technology Research and Development* **43:** 5-24.

Suthers D, Weiner A (1995) Groupware for developing critical discussion skills. In: Schnase JL, Cunnius EL *Proceedings of CSCL '95: The First International Conference on Computer Support for Collaborative Learning*. New Jersey: Lawrence Erlbaum Associates Inc., pp 341-348.

Tao PK (2001) Developing understanding through confronting varying views: the case of solving qualitative physics problems. *International Journal of Science Education* **23:** 1201-1218.

Tao PK (2003) Eliciting and developing junior secondary students' understanding of the nature of science through a peer collaboration instruction in science stories. *International Journal of Science Education* **25:** 147-171.

Williams A (1995) Long-distance collaboration: a case study of science teaching and learning. In: Spiegel SA *Perspectives from Teachers' Classrooms. Action Research. Science FEAT (Science for Early Adolescence Teachers).* Tallahassee, Florida: Southeastern Regional Vision for Education.

Zohar A, Nemet F (2002) Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching* **39:** 35-62.

Three 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:

- Keys (1997) and *Keys (1995)
- Sherman and Klein (1995a) and *Sherman and Klein (1995b)
- Tao (2001) and *Tao (2000b)

Full references for subsidiary papers (asterisked*) are given in the References section (Chapter 6) of this review. For the remainder of this chapter of the report and throughout the findings and conclusions in Chapter 5, the lead paper only is cited and any 'a' or 'b' notation is dropped.

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

Countries of studies

Table 4.1 shows the countries in which studies selected for in-depth review were carried out. The majority of the studies were undertaken in the USA, with others as detailed below. The proportion of studies undertaken in the USA (60%) is larger than the proportion of studies from the USA in the map (39%). The in-depth sample does not include research carried out in the UK. This contrasts with the 12 UK studies in the map (94 studies).

Table 4.1: Countries in which the studies selected for in-depth review were carried out (N = 10, mutually exclusive)

Country	Number of studies	Study
USA	6	Keys 1997 Lavoie, 1999 Palincsar *et al.*, 1993 Sherman and Klein, 1995a Suthers and Weiner, 1995 Williams, 1995
China: Hong Kong	2	Tao, 2001 Tao, 2003
Canada	1	Lajoie *et al.*, 2001
Israel	1	Zohar and Nemet, 2002

The researchers

Of the ten studies, half appeared to be undertaken by single researchers, of whom one (Williams, 1995) was clearly identified as a practitioner researcher, one seemingly resulted from PhD studies (Keys, 1997) and three were completed by post-doctoral or senior researchers (Lavoie, 1999; Tao, 2001, 2003). Three of the studies were undertaken by pairs of researchers (Sherman and Klein, 1995; Suthers and Weiner, 1995; Zohar and Nemet, 2002) and two by teams of three or four researchers (Lajoie *et al.*, 2001; Palincsar *et al.*, 1993).

Almost all the authors appear to be based in universities. The exception was Williams, a school-based teacher, who was involved via a university project. In a small number of cases, the researchers participated in teaching or supporting the activities for the study: for example, Keys (1997); two of the researchers (not named) in Palincsar *et al.* (1993) and Nemet in Zohar and Nemet (2002). In one study (Lavoie, 1999) the author carried out the study in collaboration with five 'teacher/researchers'.

On the basis of information provided, it appeared that three studies were externally funded: Lajoie *et al.* (2001), Palincsar *et al.* (1993), and Suthers and Weiner (1995).

Subject focus

Half of the ten studies in the in-depth review focused on small-group discussions in integrated science lessons, four in biology, one in physics and none in chemistry. This constitutes a noticeable rise in proportion of biology studies (from 19% in the map) and a sharp fall in proportion of physics studies in comparison with the 39% physics-based studies in the map. This difference may be due to the fact that understanding evidence comes to the fore in particular when discussing contentious issues, which often are related to biology: for example, genetic engineering or human immuno-deficiency virus (HIV)-acquired immuno-deficiency syndrome (AIDS). However, the low number of studies in the in-depth review may well have contributed to the difference in proportions.

While all the studies involved students in using evidence, they were based on a range of science topics and different aspects of using evidence. Two studies addressed specific areas where students encounter difficulties in understanding science ideas (Palincsar *et al.*, 1993: kinetic theory; Tao, 2001: mechanics). One had a specific focus on socio-scientific issues (Zohar and Nemet, 2002, genetic engineering). One involved predictions based on evidence presented (Lavoie, 1999: biology). Three looked primarily at scientific method (Lajoie *et al.*, 2001: confirming or refuting hypotheses on disease diagnosis; Sherman and Klein, 1995: designing controlled experiments; Williams, 1995: model building, using biological material). Reasoning and argumentation skills were of interest to Keys (1997), who investigated the use of scientific reasoning skills in collaborative report-writing; Suthers and Weiner (1995) developed scientific argumentation and reasoning, using HIV-AIDS as a case study. Tao (2003) described how science stories can be used to elicit students' understanding, or not, of the nature of science.

Ages of learners in studies

The studies were undertaken with a diversity of age ranges of learners, as summarised in Table 4.2. The 90% of studies involving junior secondary (ages 11 to 15) was close to that of all studies in the map (89%), while the percentage of studies with senior secondary students (10%) was lower than the percentage of those found in the map. (Proportions are difficult to compare as the borderline between junior and senior secondary varies between studies and with sub-samples.)

Table 4.2: Ages of learners in studies selected for in-depth review

Age range	Number of studies	Study
16–18	1	Tao, 2001
13–15	7	Keys 1997 Lajoie *et al.*, 2001 Lavoie, 1999 Sherman and Klein, 1995a Suthers and Weiner, 1995 Tao, 2003 Zohar and Nemet, 2002
11–12	2	Palincsar *et al.*, 1993 Williams, 1995

Nature of the discussion groups

Size

Most studies used groups of the same size throughout but a few varied group size at different stages. A little over half (6) of the studies involved groups made up only of pairs of students (Keys, 1997; Lajoie *et al.*, 2001; Tao, 2001; Tao, 2003; Sherman and Klein, 1995; Suthers and Weiner, 1995) compared with 31% in the map. Two studies used only groups of three or four (Palincsar *et al.*, 1993; Williams, 1995), in contrast with 63% in the map and one study also involved groups of five or six (Williams, 1995). Suthers and Weiner (1995) involved pairs and also groups of three or four, and Zohar and Nemet (2002) used pairs and larger groups of five or six. Only one study did not provide details of group size (Lavoie, 1999). It seems that the studies involving smaller group sizes are over-represented in this set of studies when compared with the map, possibly because the development of understanding of evidence involves pitching views of participants against one another, which can possibly be done more effectively in smaller groups.

Grouping strategy

How groups were formed varied somewhat, depending on the focus of the study. In one case (Lajoie, 2001), friendship groups were preferred compared with 16% in the map. However, in four studies heterogeneous groups were deliberately created (Keys, 1997; Palincsar *et al.*, 1993; Sherman and Klein, 1995a; Tao, 2001). In another study, care was taken to promote argumentation by pairing students with differing abilities (Sherman and Klein, 1995).

Four studies did not give any details of how groups were formed (Lavoie, 1999; Suthers and Weiner, 1995; Tao, 2003; Williams, 1995). The proportion of studies involving the purposeful creation of heterogeneous groups was higher (five out of ten) amongst studies in the in-depth review than in the systematic map (33%).

4.3 Further details of studies included in the in-depth review and assessment of weight of evidence

Approach

Appendix 4.1 provides summary tables of the ten 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 made in the studies, the key conclusions in relation to stimuli and understanding have been presented in the authors' own words.

This section uses the data extracted from the ten studies to provide further information about the studies and to demonstrate how judgements about the weight of evidence were made. Section 4.3.1 provides an overview of the aims of the studies. 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 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).

It was important to ensure that appropriate and consistent judgements over weights of evidence were made in the review-specific areas: that is, B and C (and therefore, ultimately D, the overall judgement, which takes into account B and C). The Review Group therefore developed a table of specific indicators for weight of evidence to be used in making judgements. These are described in section 2.3.3, and presented as a table in Appendix 2.5.

The discussion in sections 4.4 and 5.1 should be read in conjunction with the table in Appendix 2.5.

Finally, having taken cognisance of the overall (D) weightings given in section 4.3.5, the findings of those studies considered to be of sufficiently high quality to be considered in detail and discussed as findings of this review are given in Table 4.7 in section 4.3. (The findings of all ten studies can be found in the summary tables in Appendix 4.1.)

4.3.1 Overview of the studies

Aims of studies

Two particular features of the reports were apparent when considering the aims of the studies. First, a characteristic of a number of the studies is that they have a diversity of aims, not all of which relate to students' use of evidence in small-group discussion work. For example, Tao (2001) was also investigating problem-solving skills. Second, the term 'collaborative learning' was often used as an umbrella term without precise definition, but it implied that it automatically included small-group discussion work of some form.

All the studies focused on evaluation of intervention programmes which had as one of their aims the promotion of small-group discussion activities.

Of the intervention evaluations, three studies (Lajoie *et al.*, 2001; Sherman and Klein, 1995; Suthers and Weiner, 1995) focused on the effect of a specific type of discussion stimulus: that is, computer-supported learning environments (CLEs*)*. The role of the computer in the studies varied from a tool for directing discussions, recording these, or providing external data into the discussions. The main aims of these studies are as follows:

- to identify the effect of types of features of a specific CLE (related to the digestive system) on student actions and verbal dialogue, and thus pinpoint features most conducive to learning and scientific reasoning (Lajoie *et al.*, 2001)

- to investigate the effects – in terms of conceptual understanding, attitude and group behaviour – of verbal interaction cues and ability groupings within a co-operative CLE (Sherman and Klein, 1995a)

- to undertake a formative evaluation of a specific CLE to stimulate collaborative formulation of a scientific argument, and thus to promote learning of science concepts and reasoning (Suthers and Weiner, 1995)

Five of the intervention evaluations explored the effect of a range of teaching strategies involving small-group discussions. The main aims of these studies are as follows:

- to investigate the use of reasoning strategies through a collaborative report-writing task in order to generate meaningful scientific models and the evidence for improvement in students' reasoning discourse (Keys, 1997)

- to examine the effects – in terms of teacher and student attitudes and their conceptual understanding and logical thinking abilities – of including a prediction/discussion phase prior to a traditional learning cycle (exploration, term introduction, concept application) (Lavoie, 1999)

- to explore whether and how group discussion of feedback of multiple alternative solutions to qualitative physics problems helps to improve students' problem-solving skills and understanding of underlying physics concepts (Tao, 2001)

- (not very specific) to assess the benefits to students of a project (on abiotic and biotic materials) completed in collaboration with a distant school (Williams, 1995)

- to determine to what extent the use of science stories could elicit an understanding of the nature of science through peer collaboration (Tao, 2003)

Two studies evaluated an intervention with a major metacognitive component. The main aims of these studies are as follows:

- to evaluate the effects of an intervention, including guidance of the use of scientific explanations and constructive group interaction on the ability to apply knowledge of kinetic molecular theory to everyday problems (Palincsar *et al.*, 1993)

- to examine the effects of a unit, which teaches argumentation skills in the context of dilemmas in human genetics, on the development of biological understanding and argumentation skills (Zohar and Nemet, 2002)

4.3.2 Methodological considerations

Study designs

The study designs were roughly balanced between naturally-occurring evaluations (Keys, 1997; Palincsar *et al.*, 1993; Suthers and Weiner, 1995; Tao, 2001; Tao, 2003; Williams, 1995) compared with researcher-manipulated evaluations (Lajoie *et al.*, 2001; Lavoie, 1999; Sherman and Klein, 1995; Zohar and Nemet, 2002). Of the latter, only one (Sherman and Klein, 1995) used a randomised control trial (RCT) design. This balance is very similar to that in the map where 22 of the 47 evaluation studies were naturally occurring and 25 researcher-manipulated, with13% (12 out of 89) being randomised controlled trials.

It should be noted that a few of the studies included only a minor evaluative component. For instance, the study by Keys (1997) set out to document changes in conceptual knowledge of students participating in small-group discussions and the author documented this knowledge before and after the intervention through interviews. However, the main focus of the study was on the characteristics of the discourse used by students when participating in specific small-group discussions.

Sample size and sampling method

None of the studies in the in-depth review 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. All studies used a convenience sample for the identification of schools, often using schools where access has been secured through previous involvement of the researcher (for instance, Lajoie *et al.*, 2001; Tao, 2003; Suthers and Weiner, 1995), or where a researcher has been on the staff as a teacher (for instance, Williams, 1995). Such convenience sampling is probably realistic for research studies fitting in with practice.

Within schools, all studies used classes as the initial unit of sampling. The selection method of classes was mostly unspecified, or based on teachers' willingness or interest (for instance, Lavoie *et al.*, 1999; Suthers and Weiner, 1995). Almost all studies took the individual student as the unit of their evaluation. Thus they measured and reported the effect of interventions on the individual's understanding. Only the studies by Keys (1997), and Suthers and Weiner (1995), all small-scale studies, explicitly took discussion groups as the unit for which the effect of the intervention is described and evaluated. By his own admission, Tao (2001) realised that a contradiction exists in his study in this regard as the pre-intervention problem-solving skills were measured for the pairs of students, whereas the post-intervention skills were documented individually. Tao (2003) took both approaches in that he looked at the achievement of pairs in pre- and post-tests, and during observations, but interviewed students individually to probe their understanding of the nature of science more deeply and in relation to their pre- and post-test performances.

Three studies (Keys, 1997; Suthers and Weiner, 1995; Tao, 2001) worked with samples equal to, or less than, one class. A few studies used samples of between 25 and 100 students (Lajoie *et al.*, 2001; Williams, 1995), and four of the studies (Lavoie, 1999; Palincsar *et al.*, 1993; Sherman and Klein, 1995a; Zohar and Nemet, 2002) used quite sizeable samples, involving eight to ten classes. Tao (2003) used 150 students from four classes for multiple choice pre- and post-tests but concentrated on one class (36 students) for more detailed observations and interviewed a subset of 18 students.

The majority of studies provide limited information about the characteristics of students in the sample. Some samples were clearly atypical: for instance, highly motivated students (Tao, 2001), mainly from lower socio-economic backgrounds (Lavoie, 1999; Suthers and Weiner, 1995), higher ability boys (Tao, 2003), or students at a private girls' school (Lajoie *et al.*, 2001).

Only three of the studies explicitly state to what extent the findings were thought to be generalisable. Keys (1997) specifically warns against generalising her findings from her interpretive study as the effects of the intervention under scrutiny. Lavoie (1999) restricted claims to classes taught by teachers trained for, and committed to, the specific intervention. Tao (2003) points out that the findings of his study were not generalisable because the sample was atypical as it consisted of high ability and well motivated students from only one school. However, it seems many studies made an implicit claim for generalisibility: for instance, as a response to research trends (Palincsar *et al.*, 1993).

Comparison/control of independent variable

Three sizeable studies (Lavoie, 1999; Sherman and Klein, 1995; Zohar and Nemet, 2002) and one small study (Lajoie *et al.*, 2001) included a comparison group. Lavoie (1999), and Zohar and Nemet (2002) compared the learning effect for groups undergoing an intervention with small-group discussion work, with those who learn through a traditional learning sequence. Two studies compared the learning effect for groups, each using small-group discussions. However, the experimental group had undergone an intervention specifically aimed at facilitating group interaction. Lajoie *et al.* (2001) looked at the benefit of special scaffolding by the teacher, and Sherman and Klein (1995a) studied the difference between a cued and uncued software program. One study compared groups with different abilities (Sherman and Klein, 1995a), undergoing the same intervention.

Several of these studies carefully matched the experimental and control groups for teacher (Lajoie *et al.*, 2001; Lavoie, 1999; Zohar and Nemet, 2002) or students' prior conceptual understanding (Lavoie, 1999; Zohar and Nemet, 2002).

Sherman and Klein (1995a) used a 2 x 3 design for high, low and mixed ability clusters, each using a cued or non-cued version of the same CBI package.

The remaining six studies did not report the use of a comparison group. They are prospective single cohort studies, classified within EPPI-Centre terminology as naturally-occurring evaluations.

Pre-post data-collection of dependent variable

Seven studies used a prospective design and collected pre- and post-intervention data, frequently using the same instrument. Some of these instruments measured students' understanding of evidence. For instance, Zohar and Nemet (2002) measured students' argumentation skills (with an identical *and* an equivalent task in the post-test) and Lavoie (1999) their logical thinking skills. Tao's (2003) very specific focus was on aspects of the nature of science which he tested before and after the students had read and discussed four stories about scientific discoveries and developments. Several studies measured effect by pre-post testing other variables; pre-post intervention tests were used by Lavoie (1999); Palincsar *et al.* (1993), and Zohar and Nemet (2002). Pre-post intervention interviews were used by Keys (1997) for documenting changes in conceptual understanding.

Some studies seemed to document benchmark data, but these are not comparable with the outcome data. For instance, Tao (2001) reports pre-intervention problem-solving success and post-intervention levels of conceptual understanding. In his study, Tao (2003) used his pre-post data to identify pairs and individuals who had or had not made changes in their understanding and then interviewed a subset of each type.

Sherman and Klein (1995a) collected pre-intervention data on students' conceptual understanding, not for a direct comparison with students' post-intervention understanding, but as a basis for pairing students in discussion groups.

Three studies do not report benchmark data to be compared with the outcomes (Lajoie *et al.*, 2001; Suthers and Weiner, 1995; Williams, 1995).

Reliability and validity of data-collection methods and tools

Only two of the studies used existing tools. Lavoie (1999) used an existing test for procedural skills: Processes of Biological Investigation Test (PBIT) with a Kuder-Richardson reliability of 0.83, and the existing Group Assessment of Logical Thinking (GALT) test with a Cronbach alpha reliability of 0.85. Tao (2003) adopted a test of understanding of the nature of science developed and tested by a previous researcher.

Only one study established the reliability of the self-designed tests used. Sherman and Klein (1995a) developed two tests: one with multiple-choice items and one with a Likert-scale structure. The reliability of each test was reported using Kuder-Richardson (value 0.87) and Cronbach alpha methods (0.78).

Tao (2001) used tools with multiple items measuring the same concept, thus implicitly increasing the reliability but no inter-item reliability score is provided.

More detail is provided on the validity of the studies. Tao (2001) and Lavoie (1999) tested equivalence of self-designed pre- and post-tests with a Spearman-Brown split-half method using a class in another school. For instance, Tao (2001) established through a Mann-Witney test no significant differences in scores ($p = 0.87$). He concludes that the level of difficulty in pre- and post-test is the same. Lavoie (1999), and Zohar and Nemet (2002) had all items in their self-designed pre-post tests checked for content validity by an 'expert'. Tao (2003) asked five experienced science teachers to check that his intervention (science stories) were appropriate in relation to the tool (questions) to be used to judge understanding of the nature of science.

It seems that piloting data-collection instruments and strategies probably occurs more often than it is reported. No studies reported field-testing the instrument and only one reported the procedure. Keys (1997) provided an exercise for students to practice collaborative report writing. Tao (2003) piloted his science stories (the intervention) with students in another school for understanding.

Some research designs in themselves provide validity. For instance, the action research design of Suthers and Weiner (1995) uses tasks which are modified and refined for each subsequent research cycle, thus increasing the validity of the data over the lifetime of the project. In addition, Keys' (1997) interpretative study provides very detailed descriptions of the context of the school, students and data-collection situation, resulting in a very high context validity.

Several studies provide little or no detail for judging the reliability or validity of the data-collection method and tools (Palincsar *et al.*, 1993; Williams, 1995).

Reliability and validity of data-analysis methods

All but two studies (Keys, 1997; Williams, 1995) reported the use of some form of statistical analysis which, if done appropriately, provides a measure of reliability for the analysis.

Two studies (Palincsar *et al.*, 1993; Lavoie, 1999) used t-tests for identifying the significance of differences in the conceptual understanding of two subsequent cohorts undergoing slightly modified interventions (Palincsar *et al.*, 1993) and in the change in understanding of experimental and control groups after an intervention (Lavoie, 1999). They also provide details of group sizes, mean scores and standard deviations, t-values and p-values. Palincsar *et al.* (1993), for instance, used t-tests to establish that conceptual knowledge gains from small-group discussions are significantly higher (t (82) = 2.625, p = 0.005) for those students who used an open-ended, problem-solving task than for those using a more closed task.

Three studies (Lajoie, 2001; Sherman and Klein, 1995; Zohar and Nemet, 2002) used ANOVA methods for identifying the significance of differences in the performance of various groups (for instance, experimental versus control groups) after an intervention. The reports of two of these studies (Lajoie, 2001; Sherman and Klein, 1995) provide details of group sizes, mean scores and standard deviations, F-values, degrees of freedom and p-values. Sherman and Klein (1995) used ANOVA analysis to identify a number of differences between their six groups: for example, the students on the cued version of the CBI performed significantly better on the post-test than those on the non-cued version (F (1,225) = 12.97, p < 0.001). ANOVA and subsequent Tukey HSD pair analysis showed that the mean performance score for the three ability groups was also significantly different.

Sherman and Klein (1995) analysed ten Likert-scale type items on attitude towards the intervention with multi-analysis of variance (MANOVA) to show no significant differences in attitudes for ability groups or cued versus non-cued software versions.

The findings of Lajoie (2001) were based on Pearson correlation analysis.

Tao (2001) used the Wilcoxon signed rank test for pre-post scores, providing a two-tailed significant level of 0.037. He concludes a significant improvement from pre- to post-testing at p = 0.05 level.

One way of addressing the reliability of the coding or grading of the data was the use of two independent markers of written test responses. Tao (2001) used 25% of all written responses and reported a high (non-specified) inter-marker agreement. Zohar and Nemet (2002) report at least 85% inter-rater agreements for an unspecified percentage responses to the three tests they used.

Keys (1997) provides an example of a reliability check for observation data when she used blind-coding of about 10% of students’ oral reasoning strategies with initial agreement of 85%.

Triangulation was a method often used for data analysis, but its usefulness for validity is rarely highlighted by the authors. For instance, Lajoie *et al.* (2001), Sherman and Klein (1995a) and Tao (2001) collect computer logs, students’ written work and video-recorded student interaction, but none of the studies describes how the multi-sources have been integrated. On the other hand, the smaller-scale study by Keys (1997) describes triangulation for validation of assertions in detail and to great effect as does Tao (2003) when comparing detailed data from 18 interviews with findings from tests and observations.

Grounded theory has been used in three studies for developing categories of interactions and use of knowledge during group discourse (Keys, 1997; Lajoie, 2001; Palincsar *et al.*, 1993). Keys (1997) mentions that she used Kuhn's framework for reasoning strategies as a basis for her grounded theory analysis of clinical interviews, thus increasing the validity. Lajoie (2001) used data from 'experts' for determining his typology for student performance in scientific reasoning.

Tao (2003) generated tentative assertions from the interactions of collaborating pairs of students and then used these to search systematically for evidence that would confirm or refute the assertions. Some assertions were abandoned or modified and new ones generated. Testing tentative assertions against data and this 'natural history' approach was seen as a possible way of improving the credibility of the results.

An interpretive study like that by Keys (1997) would not focus on the reliability of the analysis of the data, as the intention is to provide as full a picture as possible, crystallising the information around a limited number of assertions supported by descriptive data.

Apart from a description of the statistical methods, half of the studies provide little or no detail of issues related to reliability or validity of the data analysis (Lajoie *et al.*, 2001; Palincsar *et al.*, 1993; Sherman and Klein, 1995a; Suthers and Weiner, 1995; Williams, 1995).

Weight of evidence A

Taking account of the different methodological aspects above, the quality of the ten studies can be summarised as in Table 4.3 below. 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, specifically *Taking account of all quality assessment issues, can the study findings be trusted in answering the study question(s)?*

Table 4.3: Quality of the studies (weight of evidence A)

Study	Quality of the study (weight of evidence A)
Keys, 1997	Medium-high
Lajoie *et al.*, 2001	Medium
Lavoie, 1999	Medium
Palincsar *et al.*, 1993	Medium
Sherman and Klein, 1995	High
Suthers and Weiner, 1995	Medium-low
Tao, 2001	Medium
Tao, 2003	High
Williams, 1995	Low
Zohar and Nemet, 2002	Medium

4.3.3 Appropriateness of the studies' research design for the in-depth review

This section of the report synthesises the evidence from the ten studies in terms of the appropriateness of the research design for the in-depth review question. This will provide the weight of evidence category B (weight of evidence B).

The in-depth review question is:

What is the evidence from evaluative studies of the effect of different stimuli (print materials, practical work, ICT, video/film) in small-group discussions on students' understanding of evidence in science?

Research designs are weighted according to one precondition: the evaluative component of the study needs to apply to the effect of students' understanding of evidence. In addition, five design aspects are graded: the appropriateness of the sampling, the comparison with the independent variable (small-group discussions), the prospectiveness of the dependent variable (understanding of evidence), the appropriateness of the data collection and the appropriateness of the analysis methods.

Evaluative component of studies

All studies, apart from Keys (1997), include a substantive evaluative component for students' understanding of evidence. As mentioned above, evaluation plays a minor role in Keys' descriptive study, and this evaluation focused on students' understanding of science concepts rather than evidence. Thus the appropriateness for this in-depth review is low.

Appropriateness of sample size and sampling method

Since the in-depth review intends to establish broadly generalisable evidence for the effect of small-group discussions, a sampling method aimed at representativeness strengthens the weight of evidence for the findings of a study.

All studies – apart from those by Keys (1997), Lavoie (1999) and Tao (2003) – lack detail on claims of generalisibility and the sampling methods employed.

For the purposes of this review, a sample size of over 90 students, or three classes, is considered reasonable for generalising findings and conclusions. The four largest studies (Lavoie, 1999; Palincsar *et al.*, 1993; Sherman and Klein, 1995; Zohar and Nemet, 2002) used between eight and ten classes, and are thus more appropriate for this in-depth review.

The use of discussion groups as units for evaluation reduces the validity of the study's findings for this review, as literature shows (for instance, Campbell *et al.*, 2000) that publicly negotiated meaning in groups does not always equate personal conceptual understanding. This decreases the appropriateness of three studies (Keys, 1997; Suthers and Weiner, 1995; Tao, 2001).

Although students in most studies represented a reasonable cross-section of socio-economic, cultural, ability, attitudinal and gender characteristics, the studies by Lajoie *et al.* (2001), Lavoie (1999), Tao (2001), Tao (2003), and Suthers and Weiner (1995) used atypical samples and would therefore have limited generalisibility.

Appropriateness of comparison of independent variable (i.e. small-group discussions)

The in-depth review requires a design with a control group as comparison. Only three studies (Lavoie, 1999; Sherman and Klein, 1995; Zohar and Nemet, 2002) use a control group. Two of these studies carefully matched the experimental and control groups for a teacher effect and for students' prior conceptual understanding (Lavoie, 1999; Zohar and Nemet, 2002). The other study took great care in the control of external factors when constituting their small-groups according to prescribed characteristics (Sherman and Klein, 1995).

Appropriateness of data collection of dependent variable (i.e. understanding of evidence)

Studies with a prospective design measuring students' understanding of evidence before and after an intervention are most appropriate to this in-depth review. This applies to studies by Lavoie (1999), Tao (2003), and Zohar and Nemet (2002). Other studies did not collect any benchmark data or used pre-post intervention measures to establish change in students' conceptual understanding.

Appropriateness of addressing issues of reliability and validity in data collection

Section 4.3.2 summarises ways in which issues of reliability and validity of the data-collection methods and tools are addressed for each study as a whole. In general, these descriptions are equally relevant for the in-depth review question. The use of the well-established GALT test for logical thinking by Lavoie (1999) is particularly relevant for the effect of small-group discussions on students' understanding of evidence, and so is the reliability check, using the Kuder-Richardson method for the self-designed test by Sherman and Klein (1995) and the test devised by Solomon (1996) and used by Tao (2003).

The check by an external expert of the content validity of the instruments for measuring the understanding of evidence used by Lavoie (1999), Tao (2003) and

Zohar and Nemet (2002) is worth mentioning. One study (Keys, 1997) used a pilot in order to increase validity of the instruments or the data-collection strategy.

Appropriateness of addressing issues of reliability and validity in data analysis

Section 4.3.2 summarises ways in which issues of reliability and validity of the data-analysis methods are addressed for each study as a whole. In general, these descriptions are equally relevant for the in-depth review question. However, several of the elaborate statistical analysis methods focus on effects other than students' understanding of evidence. Usually they measure the effects on students' conceptual understanding or their attitudes. Some t-tests (Lavoie, 1999) and several of the ANOVA methods (Sherman and Klein, 1995; Zohar and Nemet, 2002) help ensure validity and reliability, as does Tao's (2003) system of using three methods (tests, observations and interviews) that are supportive of his 'interpretive approach'.

Weight of evidence B

Taking account of the different methodological aspects above, the quality of the ten studies can be summarised as in Table 4.4. These quality weightings have been made against the appropriateness of the study design for the in-depth review question.

Table 4.4: Appropriateness of the study design (weight of evidence B)

Study	Appropriateness of the study design for the in-depth review question (weight of evidence B)
Keys, 1997	Medium–low
Lajoie *et al.*, 2001	Low
Lavoie, 1999	Medium-high
Palincsar *et al.*, 1993	Medium-low
Sherman and Klein, 1995	Medium
Suthers and Weiner, 1995	Low
Tao, 2001	Low
Tao, 2003	Medium-high
Williams, 1995	Low
Zohar and Nemet, 2002	Medium-high

4.3.4 Relevance of the studies' focus for the in-depth review

Further features of the study designs are selected for their appropriateness for the in-depth review question. Aspects of the way in which the variables are formulated and explicated are selected for determining the relevance of the study's focus. These five aspects are discussed in this section and will each contribute to the weight of evidence for category C.

The relevance of the focus of the ten studies will be weighted according to five aspects: the representativeness of the small-group discussions; the extent to which the independent variable (stimulus) is the focus of the study; the nature and breadth of the measures for the dependent variable (understanding evidence); and the representativeness of the learning context.

Representitativeness of the stimulus

Three types of stimuli were used to promote discourse in these ten studies: printed materials, ICT and practical work. In some studies, the situation was a little more complex in that the principal stimulus was in the context of another stimulus. In the Keys (1997) study, for example, the central stimulus for the small-group discussion was written prompts for collaborative report writing in the context of laboratory work.

Five studies were based on printed materials and most used these materials in a representative manner. These included Lavoie (1999), in which written material was employed asking for written predictions, and Zohar and Nemet (2002), in which written work sheets were used as a basis for developing argumentation skills. The studies that were not quite so typical were Keys (1997), in which students wrote up collaborative laboratory reports from written questions or prompts, and the two Tao studies. In Tao (2001), the students were presented with multiple solutions to physics problems and had to resolve these through discussion. In the second Tao (2003) study, science stories were used to evaluate student understanding of the nature of science. In no study was the content of written material used in a totally untypical way.

Five of the studies involved ICT and four used the software in a representative manner to teach aspects of science, in particular the use of evidence. A good example is Lajoie *et al.*'s study (2001) in which the BioWorld programme taught students the use of evidence that allowed them to request items of information in an interactive way in order to diagnose diseases. Another representative use of ICT was in the study by Sherman and Klein (1995), in which students were offered verbal interaction cues to facilitate summarising and explaining between pairs of students. However, the main purpose of the Suthers and Weiner (1995) study was software development.

The practical work described in the Palincsar *et al.* (1993) study was very typical of work on coloured solutions carried out in school science laboratories. The practical work described by Williams (1995), on the other hand, was rather atypical. The aim of this study was for the students to exchange the items collected in their school environment with items collected by students in a distant school and use these for interpretive ecological work.

Focus of the stimuli

As was the case with a number of the studies in the map, some of the ten studies had multiple aims and the factor of interest to this review (i.e. nature of the stimulus used) was not the sole part of the intervention; alternatively, more than one stimulus was used and the effects of these were not separated.

Those studies in which the stimulus was the sole and explicit independent variable were Lajoie *et al*. (2001), Sherman and Klein (1995), Suthers and Weiner (1995), Tao (2001), Tao (2003) and Williams (1995).

Studies which had more than one stimulus in the intervention are as follows:

- Lavoie (1999) – written material and a prediction task before the main discussion task
- Palincsar *et al.* (1993) – practical work supported by instruction and guidance in the use of scientific explanations and social norms conducive to constructive interaction
- Zohar and Nemet (2002) – written materials and training in argumentation

In the study by Keys (1997), the stimulus was only a small part of a more complex intervention.

Those studies which specifically evaluated individual variables are the most useful for the purposes of this review.

The nature and breadth of the dependent variable (i.e. the understanding of evidence)

As mentioned previously, understanding of evidence as defined for this review has three aspects. In order of progressive sophistication, it involves engaging with primary or secondary data; second, it requires developing models or claims; and, third, it allows drawing on data to justify models, claims or arguments.

One of the studies (Lavoie *et al.*, 1999) has a main focus on engagement with data. This looks at the effect of making sense of individual predictions in small-group discussions on conceptual understanding.

Two studies (Keys, 1997; Palincsar *et al.*, 1993) focus on the role of explanations in constructing conceptual models or claims from experimental or print data: Palincsar *et al.* (1993) explore the role of scaffolding the explanation process in model construction; Keys (1997) looks at model reconstruction in the light of a variety of sources of information. Keys identifies different aspects of the understanding of evidence, such as the ability to recognise that a current model may be incorrect; to generate new hypotheses and test these; to evaluate new data for consistency with the model; and to co-ordinate data in a coherent body to support a model.

Three studies (Lajoie *et al.* 2001; Suthers and Weiner, 1995; Zohar and Nemet, 2002) specifically focus on students' abilities to generate and support an argument, the highest level of understanding evidence. The studies by Lajoie *et al.* (2001), and Suthers and Weiner (1995) explore how different features of software packages may help to direct argumentation skills. Zohar and Nemet (2002) explore how this ability may be strengthened through a more complex teaching intervention.

The nature of the understanding of evidence in the studies by Sherman and Klein (1995), Tao (2001) and Williams (1995) is much more obscure. The first study measures scientific reasoning skills and process skills, while the second is interested in the change in problem-solving skills.

Tao's (2003) study is a little different in that it does not present the students with primary data to consider. Rather it elicits their understanding of the nature of science through peer collaboration about stories based on four scientific discoveries. The study focuses on scientific method and the nature of scientific theory.

The representativeness of the research context

Most studies collect data in intact classrooms (Keys, 1997; Lajoie *et al.*, 2001; Lavoie, 1999; Palincsar *et al.*, 1993; Tao, 2001; Tao, 2003; Williams, 1995; Zohar and Nemet, 2002). This research context facilitates generalisation of the findings. One approach that could compromise the generalisibility of studies' findings is found in several studies which use unnatural experimental situations: Suthers and Weiner (1995), for instance, use pairs of orally interacting students on different computers instead of one pair per computer; and Sherman and Klein (1995) asked clusters of their dyads to work outside normal classes in a special laboratory.

Weight of evidence C

Taking account of the different aspects above, the quality of the ten studies can be summarised as in Table 4.5. These quality weightings have been made against the relevance of the focus of the study for the in-depth review question.

Table 4.5: Relevance of focus of the studies (weight of evidence C)

Study	Relevance of the focus of the study for the in-depth review (weight of evidence C)
Keys, 1997	Medium
Lajoie *et al.*, 2001	Medium-high
Lavoie, 1999	Medium
Palincsar *et al.*, 1993	Medium
Sherman and Klein, 1995a	Medium
Suthers and Weiner, 1995	Medium
Tao, 2001	Medium
Tao 2003	Medium-high
Williams, 1995	Medium-low
Zohar and Nemet, 2002	Medium-high

4.3.5 Overall weighting: D

Studies were given a rating on a five-point scale in each of the categories of weight of evidence: that is, the quality of the study (weight of evidence A), the appropriateness of the study's design for this specific in-depth review question (weight of evidence B), and the relevance of the focus of the study for this in-depth review question (weight of evidence C). These weights of evidence, together with the overall weight for each study (weight of evidence D), are summarised in Table 4.6. The points on the scale are as follows:

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

Table 4.6: Weights of evidence assigned to studies

Study	Weight of evidence A	Weight of evidence B	Weight of evidence C	Weight of evidence D
Keys, 1997	MH	ML	M	M
Lajoie *et al.*, 2001	M	L	MH	M
Lavoie, 1999	M	MH	M	M
Palincsar *et al.*, 1993	M	ML	M	M
Sherman and Klein, 1995	H	M	M	MH
Suthers and Weiner, 1995	ML	L	M	ML
Tao, 2001	M	L	M	ML
Tao, 2003	H	MH	MH	MH
Williams, 1995	L	L	ML	L
Zohar and Nemet, 2002	M	MH	MH	MH

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

4.4 Synthesis of evidence

Short summaries of the relevant results from each of the studies that were judged to be of sufficient quality are given in Table 4.7. Table 4.8 shows the stimuli investigated in these seven studies. The findings from the studies were clustered according to common features agreed by two members of the team. The emerging themes are described and discussed below as the findings of this review. More detail for all ten studies considered in the in-depth review are set out in Appendix 4.1. In view of the narrow evidence base, many of the findings have been cast in tentative terms (Table 4.7).

Table 4.7: Summary of results of studies on students' understanding of evidence.

Paper	Results: students' understanding of evidence	Weight of evidence D
Keys, 1997	Types of scientific reasoning and method used by students • Recognising that prior ideas (models) may be incorrect • Evaluating new observations for consistency with current ideas and using evidence to modify ideas • Co-ordinating all mutually consistent knowledge propositions into a coherent model	Medium
Lajoie *et al.*, 2001	Types of scientific reasoning and method used by students • Argumentation leading to hypothesis formation • Collection of relevant evidence to confirm or refute hypothesis • Revision of initial hypothesis Impact of adult guidance	Medium

Paper	Results: students' understanding of evidence	Weight of evidence D
	• Students working without adult support spent more time at the beginning on insignificant details but benefited from generating their own hypotheses, and followed up on their own problem-solving strategies.	
Lavoie, 1999	• Prediction/discussion-based learning cycle (hypothetico-predictive reasoning) instruction compared with traditional learning cycle instruction produced significant gains in the use of process and logical thinking skills, science concepts and scientific attitudes. • In general, teachers felt that the learning cycle instruction was more effective than their normal teaching mode for revealing students' misconceptions, teaching process skills and teaching some concepts. • Student questionnaire data revealed strong trends favouring learning cycle instruction.	Medium
Palincsar *et al.*, 1993	Types of scientific reasoning and method used by students • Contradictions in results led to discussion of accuracy of method and of reporting. • The use of explanations to scaffold discussions, particularly to provide reasons for proposals, aided student discourse. • Use of previous information to inform planning of investigations • Students demonstrated some of the characteristics of engaging in the enterprise and language of science, particularly in the second year of the study. • The open-ended investigation task instigated in year two of the study developed more complex understanding of evidence than the closed practical tasks employed in year one.	Medium
Sherman and Klein, 1995	• Students using the cued version of the computer program performed significantly better on the post-test than students using the non-cued version. • Direct observation of students showed that students in cued dyads exhibited significantly more summarising and helping behaviours than non-cued students. • Higher ability dyads exhibited significantly less off-task behaviour than the other dyads.	Medium-high
Tao, 2003	• Many students have entrenched inadequate views on the nature of science (NOS). • Students can give articulate and sophisticated arguments, irrespective of whether these views are adequate or not. They draw on prior knowledge and/or science stories for such arguments. • The science stories on the NOS instruction influence students in substantial ways but not always to improve understanding. • When studying the science stories many students selectively attend to certain aspects that appear to confirm their inadequate views of NOS. • The peer collaboration provided students with experiences of conflict and co-construction that helped them develop shared understanding of NOS. However, many students interpreted	Medium-high

Paper	Results: students' understanding of evidence	Weight of evidence D
	the science stories in idiosyncratic ways other than as intended by the instruction and subsequently changed from one set of inadequate views of NOS to another rather than to adequate views.	
Zohar and Nemet, 2002	• Following instruction in a unit that teaches argumentation skills, the number of students using correct, specific biological knowledge in constructing arguments trebled. • Students in the experimental group scored significantly higher than students in the control group in a test of genetics knowledge. • Analysis of the written tasks showed an increase in the number of justifications and in the complexity of argument. • Students were able to transfer reasoning abilities tools in the context of bioethical dilemmas to the context of dilemmas taken from everyday life. • There were dramatic changes in the quality of student arguments. • Changes were detected in the frequency of explicit conclusions, the mean number of justifications for a conclusion and in the number of ideas students expressed while talking. • Integrating explicit teaching of argumentation into the teaching of dilemmas in human genetics enhances performance in both biological knowledge and argumentation.	Medium-high

Table 4.8: Relevant medium-high and medium quality studies and the stimuli they investigate

Study	Stimulus	Understanding of evidence in science
Keys, 1997	Printed materials – structured eport guidelines	Pairs of students writing collaborative reports of practical work
Lajoie *et al*., 2001	Computer-based learning environment to diagnose diseases	The process of scientific reasoning adopted by pairs of students when using evidence to tests hypotheses
Lavoie, 1999	Print materials	To encourage hypothesising and prediction in small groups
Palincsar *et al.*, 1993	Practical work (Chemistry)	The use of scientific explanations and social norms conducive to constructive interaction in groups of 3 to 4
Sherman and Klein, 1995	Computer-based	Collaboration by pairs of students in scientific method and designing controlled experiments
Tao, 2003	Print materials - science stories	Eliciting discussion with pairs of students about the nature of science and how scientists work
Zohar and Nemet, 2002	Print materials – Genetic Revolution Unit	Argumentation promoted through biological evidence used by small groups when discussing dilemmas

The review shows that a variety of stimuli (printed curriculum material, computer software and practical work) can support students' understanding of evidence/the nature of science through small-group discussions. This is demonstrated by the positive results from the studies using:

- computer software (Lajoie *et al.*, 2001; Sherman and Klein, 1995)
- printed curriculum materials (Keys, 1997; Lavoie, 1999; Zohar and Nemet, 2002)
- practical work (Palincsar *et al.,* 1993)

In addition, an intervention in one study (Tao, 2003) had variable success but nonetheless provides valuable information for this review.

All these studies present interesting insights into certain aspects of learning in their particular circumstances. These point to features of the stimuli and of the small-group discussions that need to be considered when using stimuli to enhance the understanding of evidence in science, including scientific method and the nature of science.

Due to the low number of suitable studies and their diversity, no attempt has been made to make relative judgements about the different types of stimuli. Rather, this review has discussed the features of different stimuli that do (or sometimes do not) provide students with the experience of working in a scientific way, especially in collaboration and the use of evidence.

(a) Importance of guidance for small-group discussion

The inclusion of certain types of guidance in small-group discussion, irrespective of the nature of the stimulus, is one of the most striking factors that aids successful small-group discussion in respect of understanding evidence and the nature of science. Guidance is provided for better collaboration, better understanding or improved metacognitive strategies. Guidance can be on the substance of argumentation or on the procedure.

Four studies show that guidance can be provided in a variety of ways and two describe how lack of guidance can lead to poor or limited progress in understanding. One further study demonstrates that guidance can produce metacognition about collaboration, but not result in the actual application of collaborative reasoning.

Provide a 'checklist' for discussion

Zohar and Nemet (2002) devoted one lesson entirely to explicit instruction about argumentation during which arguments were defined and their structure explained. Good arguments, for example, include true, reliable and multiple justifications, and also refer to alternative explanations that rebut them. Students then practised the principles through several concrete examples (print materials).

The Palincsar (1993) intervention included instruction and guidance in the use of scientific explanation by way of the components of good explanation conducive to constructive interaction. The mode of discourse encouraged included identification of the chemicals involved in the practical problem and explanation of what was happening during the practical work using observations, evidence and background knowledge.

Building in an opportunity for conflict, confrontation or debate

Lavoie (1999) added a written predictive/discussion phase at the beginning of a three-phase learning cycle involving exploration, term introduction and concept application. As the students had to prepare their predictions individually, this presented opportunities for interactive debate when they then worked in small groups.

Lack of guidance limits understanding

Tao (2003) found that students working in pairs to understand the nature of science stories about scientific discoveries showed variable progress. Some pairs did show enhanced understanding when tested post intervention but other pairs made no progress and some even regressed compared with their pre-test results. This was attributed to students interpreting the stories in idiosyncratic ways other than those intended and focusing their attention selectively on certain aspects of the stories that appeared to confirm and reinforce their inadequate views.

Lajoie *et al.* (2001) used software to provide an opportunity for students to use their knowledge of biology to diagnose diseases with a view to examining how scaffolding influenced student actions. All students could use the structured nature of the programme to call for evidence and generate hypotheses through discussion. However those who worked without adult guidance were more likely to spend time on insignificant detail. In contrast, the benefit of working solely with the software was that students generated their own hypotheses and followed up on their own problem- solving strategies.

Guidance on protocol/procedure

The experimental group of students in the Sherman and Klein (1995) study, on designing controlled experiments, was allocated roles, such as summariser explainer or listener, by the computer. Built-in cues in the experimental software then facilitated discussion by prompting and thus facilitating summarising and explaining behaviour between partners. The control group had no procedural cues. As well as aiding discussion, the provision of cues improved scientific understanding.

In the Palincsar *et al.* (1993) study, students were given instruction on the social norms for discourse. These covered four aspects: contributing to the group and helping others; supporting ideas with reasons; working to understand others' ideas; and building on one another's ideas.

(b) Function of the stimulus in creating productive argumentation

Some studies consider the role of conflict in promoting productive argumentation as logical connections and that explanations can be improved when students engage in challenging and critiquing each others' ideas (Keys, 1997). Therefore, an appropriate goal of instruction is to engage all students in productive argumentation as distinct from general discussion. However, this is not a common feature of traditional science education. Five of the seven studies provided stimuli which set out to maximise debate or set up conflict situations in two different ways. Constructive debate was promoted by internal means where a diversity of views and/or understanding was represented from individual positions within a group.

Table 4.9 summarises the ways in which debate was promoted or enhanced by the stimuli in the seven studies.

Table 4.9: Summary of external and internal motivators of conflict or debate promoted by different stimuli

Nature of the conflict	**Nature of the stimulus**		
	Written	***Computer***	***Practical***
External and internal conflict	Keys, 1997 Lavoie, 1999	-	Palincsar *et al.*, 1993
External conflict only	Zohar and Nemet, 2002	-	-
Internal conflict only	-	Sherman and Klein, 1995	-
No explicit conflict	Tao, 2003	Lajoie *et al.*, 2001	-

Lavoie (1999) also used the technique of asking students to make individual predictions and then agree a joint prediction. He found that this approach seemed to promote inter-peer discussion, the use and development of the logical thinking process (by making the students' prior beliefs more explicit), and increased students' cognitive commitment. In the Palincsar (1993) study, external conflict derived from the differing results students produced in their practical work when given a task to complete in their small groups. They then had to describe and debate the results in terms of scientific explanation and in relation to the importance of the care with the practical method and accuracy of reporting. Their differing explanations provided the internally derived conflict.

(c) Nature of the task stimulus

A number of studies related the diversity and depth of the small-group discussions dealing with evidence to the complexity of the science task. Palincsar *et al*. (1993) report that an open-ended investigative task developed more complex understanding of evidence than a closed practical task. Similarly, Keys (1997) found that scientific reasoning about evidence in any practical task requires recognition of the tentative nature and existing conceptual models, and the identification of relevant observations to use as evidence for modifying these models. Domains in practical tasks allowed co-ordination of relevant conceptual propositions with the judgement of laboratory evidence for constructing new models. Equally, Lavoie (1999) showed that introducing a prediction/discussion phase before the exploration phase of the learning cycle enriched the students' experience of scientific process and resulted in improved reasoning skills, conceptual achievement and scientific attitudes.

(d) Impact of students' prior science ideas, concepts or models

One established factor which can contribute to students' learning any topic is the extent of their prior knowledge. The negative influence of this was very strong in Tao's (2003) study which investigated whether and how students develop understanding of the nature of science (NOS) through reading stories of scientific discoveries. The principal interpretation of this work was that the limited success of the NOS instruction may be attributed to the deeprootedness of students' inadequate views of NOS. (It was, however, also pointed out that alternative explanations might be the short duration of the instruction or might be due to the ways in which students make sense of these stories rather differently from those intended by the instruction.) Students interpreted the stories in idiosyncratic ways other than those intended by the instruction and focused their attention selectively on certain aspects of the stories that appeared to confirm and reinforce their inadequate views. The suggestion was made that misunderstandings about the nature of science could be avoided if there is teacher input to enhance the peer-collaboration strategy (see (a) above concerning the need to provide guidance).

The influence of prior knowledge is supported by the work of several other studies which also drew attention to this feature. Keys (1997) points out that reasoning includes the aspect of modifying prior knowledge and that small-group discussion offers students the opportunity for generalised reasoning in science learning. The types of reasoning involved include recognising that one's own prior science ideas or models may be incorrect; evaluating new observations for consistency with currently held ideas; using evidence to modify ideas; and co-ordinating all mutually consistent knowledge propositions into a coherent model.

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 2.5) worked well in securing coherence of these judgements across data-extraction teams. In addition, all four core group members 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. In order to increase the discrimination between studies, the weighting of two aspects of the algorithm has been modified slightly.

4.6 Involvement of users in the review

When the Review Group met to discuss the draft of the first report, there was a steer from members at the meeting that it would be valuable to illuminate aspects of small-group discussion work, and that a focus on the nature of the stimulus was seen as important. One head of science stressed the importance of knowing more about the variety and role of stimuli in promoting discussion.

All members of the group participated in (i) determining appropriate inclusion and exclusion criteria and in (ii) refining the review-specific keywords. Review Group members also contributed to the work by suggesting relevant studies for inclusion in the review. The study has been discussed at seminars and presentations, and feedback from researchers and practitioners has also proved valuable.

Due to the short timeline for this study, it was not practical to involve users at every stage and so not all members of the Review Group have been able to comment on this review report. Similarly, although teacher members expressed a willingness to ascertain the views of school students on the findings of the report, there has not yet been the opportunity for this to occur.

5. FINDINGS AND IMPLICATIONS

5.1 Summary of principal findings

5.1.1 Identification of studies

The overall research review question for this review is:

How are small-group discussions used in science teaching with students aged 11-18, and what are their effects on students' understanding in science or attitude to science?

Within this, the research review question identified for the in-depth review is:

What is the evidence from evaluative studies of the effect of using different stimuli (print materials, practical work, ICT, video/film) in small-group discussions on students' understanding of evidence in science?

5.1.2 Mapping of all included studies

Ninety-four 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 small-group discussions, as summarised below.

- The majority of the studies reported work that has taken place in the USA, the UK and Canada.
- Small-group discussions were used with all ages of student in the secondary age range.
- Most studies were carried out with mixed ability and mixed gender classes.
- The majority of work focused on small-group discussions in relation to students' understanding.
- The most common stimuli used to promote discussion were prepared curriculum materials followed by practical work and then computer software.
- A diversity of measures was used to assess effects on understanding and attitude.
- Very little research has been done on small-group discussions in relation to the teaching of chemistry.
- Typical small-group discussions involved groups of three to four students emerging from friendship ties, and have a duration of at least 30 minutes.

- Typical small-group discussions had individual sense-making as their main aim (as opposed to, for example, leading to a group presentation).
- The most common research strategy was that of the case study.
- Twenty-eight studies had experimental designs, of which 12 were RCTs.
- The most popular techniques for gathering data were observation, videotapes and audiotapes of discussions, interviews, questionnaires and test results.

5.1.3 Nature of studies selected for the in-depth review

Ten 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

Overall weight of evidence	Number of studies	Study
High		–
Medium-high	3	Sherman and Klein, 1995 Tao, 2003 Zohar and Nemet, 2002
Medium	4	Keys, 1997 Lajoie *et al.*, 2001 Lavoie, 1999 Palincsar *et al.*, 1993
Medium-low	2	Suthers and Weiner, 1995 Tao, 2001
Low	1	Williams, 1995

5.1.4 Summary of findings from the in-depth review

Two findings emerge most strongly from this review:

- First, small-group discussion, focused on understanding the use of evidence regardless of the prompt stimulus, is enhanced and focused by giving students some form of guidance on how to use that stimulus effectively. This guidance can be prior training in argumentation that provides instruction on how to use evidence or can be built into the structure or sequence of a stimulus-based task.
- Second, a successful stimulus for students working in small groups to enhance their understanding of evidence has two elements. One requires students to generate their individual prediction, model or hypothesis which they then debate in their small group (internally driven conflict or debate). The second element requires them to test, compare, revise or develop that jointly with further data provided (externally driven conflict or debate).

Other findings which are of interest are as follows:

- Prior knowledge can sometimes limit the understanding of evidence and its function. This can, for example, be the use of incorrect or inadequate factual knowledge or an idiosyncratic or inconsistent use of evidence to develop a hypothesis or test a model.

- Rich stimuli, such as those that provide complex and open-ended engagement, enhance opportunities for developing understanding of evidence.

Links with other reviews

This review has links with one other undertaken by the EPPI-Centre Science Review Group:

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.

The above review was based on 14 studies. Nine from that review are included in the ten studies reviewed here, together with the recently published Tao (2003). Thus the source material for the two reviews is very close. However, the ranking of the papers is somewhat different to take into account the differing objectives of the two reviews. Nonetheless, there was considerable overlap with the findings with respect to:

- the impact of guidance;
- the role of conflict.

This review has not included details about:

- different types of learner (*learner-as-explorer* and *learner-as-student*, Hogan (1999); or
- different types of understanding – science domain, science process and metacognition (Finkel, 1996).

The additional material this review contains from the Tao (2003) paper focused on the impact of prior knowledge that sometimes limits understanding of the nature of science.

5.2 Strengths and limitations of this systematic review

Strengths

The review has the following strengths:

- The review focus is highly topical. The Review Group has already been contacted by potential users interested in the findings. Further evidence of the topicality comes from the range of countries in which studies have been undertaken and from the dramatic rise in relevant published papers since 1992 as demonstrated in the map for this review (Table 3.1).

- The review has served to establish that there is consistency in the research approaches that those working in the area feel are appropriate to researching practice related to the use of small-group discussions. Such approaches make use of quantitative data, but also draw extensively on qualitative data in the form of students' written responses, interviews and audiotapes of dialogue during discussions.

- End-users of the review findings have been closely involved at all stages of the review.

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

Limitations

The review has two main limitations:

- There is a scarcity of studies that focused on the stimulus as a discrete independent variable, which resulted in very little work emerging which related specifically to the in-depth review question. Only seven studies were judged to be of reasonable quality with respect to the review question.

- 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, there was considerable variety in the nature and purpose of the discussion tasks, in the data collected, and in the interpretation of the terms *evidence* and *understanding of evidence*. Thus, teasing out the findings which specifically related to small-group discussions and to particular stimuli was not easy, and 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.

Additionally, the Review Group feel some concern about another aspect of this review. Three studies were graded as overall medium-low or low quality (category D). These were included in the in-depth review as judgements of quality are not made until comparatively late in the review process. However, this unproductive effort is a function of the process itself, rather than this specific 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 Implications for policy

The review has *not* yielded any evidence that inclusion of any specific stimulus for small-group discussions adversely affect students' understanding of the nature of evidence. However, it should also be noted that there is a scarcity of high quality research evidence in the area on which the in-depth review focused.

Where small-group discussions are advocated as a teaching approach, it is important to support this with guidance on running such discussions in a way which will increase the effectiveness of students' learning. Such guidance should include advice to students on how to use materials for the purposes of discussion, as well as the stimulus materials themselves.

5.3.2 Implications for practice

The review has indicated that there is a diversity of ways in which the term *understanding of evidence* is being interpreted. One implication for practice is therefore that teachers should be aware of this lack of clarity.

A further implication is that the success of small-group discussion, whatever the stimulus, depends in part on the students receiving some guidance on how to carry out or structure their discussions. That guidance might be written instruction, cues built into computer software or verbal support from teachers.

It can be beneficial to present a task that offers opportunities for students to generate both their own input (e.g. own predictions, hypotheses) (internal debate/conflict) and a requirement to use that in conjunction with the stimulus provided by the teacher whether written, computer software or practical work (external debate/conflict).

Tasks which are rich (i.e. complex and open-ended) are more likely to promote discussion and understanding of evidence in science than are simple or closed tasks.

Students' lack of sufficient factual knowledge of the subject of the task and/or of a systematic and consistent approach to the use of evidence can impede learning about evidence, unless support is given.

Teachers should also be aware of the lack of high quality research evidence in the area on which the in-depth review focused.

5.3.3 Research

Secondary research

Exploration of additional areas of the systematic map would appear to be particularly helpful to provide a broader picture of research findings on small-group discussion work. Such areas would include the following:

- the use of small-group discussions in relation to the development of understanding of socio-scientific issues
- aspects to do with group composition, exploring, for example, relationships between group size or gender balance within groups and development of conceptual understanding
- the effectiveness of small-group discussions for different learning outcomes (e.g. attitude, decision-making)
- the nature of small-group discussions

Primary research

One particularly strong feature which has emerged from the work undertaken for this review and the previous one (Bennett *et al*., 2004) is that there is a dearth of primary research on small-group discussion work and considerable uncertainty on the part of teachers as to what they are required to do. Both these factors point to a pressing need for a medium- to large-scale research study which focuses on the use and effects of a limited number of carefully structured small-group discussion tasks aimed at developing various aspects of students' understanding of evidence. Such research would also look at the effects of particular stimuli.

6. REFERENCES

6.1 Studies included in map and synthesis

The 94 studies included in the systematic map were reported in 119 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.*

Alexopoulou E, Driver R (1996) Small-group discussion in physics: peer interaction modes in pairs and fours. *Journal of Research in Science Teaching* **33:** 1099–1114.

*Alexopoulou E, Driver R (1997) Gender differences in small group discussion in physics. *International Journal of Science Education* **19:** 393–406.

Arvaja M, Haekkinen P, Etelaepelto A, Rasku-Puttonen H (2000) Collaborative processes during report writing of science learning project: the nature of discourse as a function of task requirements. *European Journal of Psychology of Education* **15:** 455–466.

Bianchini JA (1997) Where knowledge construction, equity and context intersect: student learning of science in small groups. *Journal of Research in Science Teaching* **34:** 1039–1065.

*Bianchini JA (1999) From here to equity: the influence of status on student access to and understanding of science. *Science Education* **83:** 577–601.

Chan CKK (2001) Peer collaboration and discourse patterns in learning from incompatible information. *Instructional Science* **29:** 443–479.

Chang CY, Mao SL (1999a) Comparison of Taiwan science students' outcomes with inquiry-group versus traditional instruction. *Journal of Educational Research* **92:** 340–346.

Chang CY, Mao SL (1999b) The effects on students' cognitive achievement when using the cooperative learning method in earth science classrooms. *School Science and Mathematics* **99:** 374–379.

Chang HP, Lederman NG (1994) The effects of levels of cooperation within physical science laboratory groups on physical science achievement. *Journal of Research in Science Teaching* **31:** 167–181.

De Vries E, Lund K, Baker M (2002) Computer-mediated epistemic dialogue: explanation and argumentation as vehicles for understanding scientific notions. *Journal of the Learning Sciences* **11:** 63–103.

Fawns R, Salder J (1996) Managing students' learning in classrooms: reframing classroom research. *Research in Science Education* **26:** 205–217.

Finkel EA (1996) Making sense of genetics: students' knowledge use during problem solving in a high school genetics class. *Journal of Research in Science Teaching* **33:** 345–368.

Ford CE (1999) Collaborative construction of task activity: coordinating multiple resources in a high school physics lab. *Research on Language and Social Interaction* **32:** 369–408.

Gayford C (1993) Discussion-based group work related to environmental issues in science classes with 15-year-old pupils in England. *International Journal of Science Education* **15:** 521–529.

Gayford C (1995) Science education and sustainability: a case-study in discussion-based learning. *Research in Science and Technological Education* **13:** 135–145.

Gilbert JK, Pope ML (1986) Small group discussions about conceptions in science: a case study. *Research in Science and Technological Education* **4:** 61–76.

Goldman SV (1996) Mediating microworlds: collaboration on high school science activities. In: Koschmann T (ed.) *CSCL: Theory and Practice of an Emerging Paradigm. Computers, Cognition and Work*. New Jersey: Lawrence Erlbaum Associates Inc.

Hogan K (1999a) Sociocognitive roles in science group discourse. *International Journal of Science Education* **21:** 855–882.

Hogan K (1999b) Thinking aloud together: a test of an intervention to foster students' collaborative scientific reasoning. *Journal of Research in Science Teaching* **36:** 1085–1109.

*Hogan K (1999c) Assessing depth of sociocognitive processing in peer groups' science discussions. *Research in Science Education* **29:** 457–477.

*Hogan K (1999d) Relating students' personal frameworks for science learning to their cognition in collaborative contexts. *Science Education* **83:** 1–32.

Hogan K (2002) Small groups' ecological reasoning while making an environmental management decision. *Journal of Research in Science Teaching* **39:** 341–368.

*Hogan K, Nastasi BK, Pressley M (2000) Discourse patterns and collaborative scientific reasoning in peer and teacher-guided discussions. *Cognition and Instruction* **17:** 379–432.

Hornsey M, Horsfield J (1982) Pupils' discussion in science: a strategem to enhance quantity and quality. *School Science Review (Science Education Notes)* **63:** 763–767.

*Howe C, Tolmie A, Anderson A (1991) Information technology and group work in physics. *Journal of Computer Assisted Learning* **7:** 133–143.

Hynd CR, McWhorter JY, Phares VL, Suttles CW (1994) The role of instructional variables in conceptual change in high school physics topics. *Journal of Research in Science Teaching* **31:** 933–946.

Jimenez-Aleixandre MP (2002) Knowledge producers or knowledge consumers? Argumentation and decision making about environmental management. *International Journal of Science Education* **24:** 1171–1190.

*Jimenez-Aleixandre MP, Bugallo-Rodriguez A (1997) Argument in high school genetics. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching. Chicago, IL, USA: March 20–24.

Jimenez-Aleixandre MP, Diaz de Bustamante J, Duschl RA (1998) Scientific culture and school culture: Epistemic and procedural components. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching. San Diego, CA, USA: April 19–22.

Jimenez-Aleixandre MP, Rodriguez AB, Duschl RA (2000a) ‘Doing the lesson’ or ‘doing science’: argument in high school genetics. *Science and Education* **84:** 757–792.

*Jimenez-Aleixandre MP, Pereiro-Munoz C, Aznar-Cuadrado V (2000b) Expertise, argumentation and scientific practice: A case study about environmental education in the 11th grade. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching. New Orleans, LA, USA: April 28–May 1.

Johnson SK, Stewart J (2002) Revising and assessing explanatory models in a high school genetics class: a comparison of unsuccessful and successful performance. *Science and Education* **86:** 463–480.

Johnston K, Scott P (1991) Diagnostic teaching in the classroom: teaching/learning strategies to promote development in understanding about conservation of mass on dissolving. *Research in Science and Technological Education* **9:** 193–212.

*Kelly GJ, Crawford T (1995) Computer representations in students’ conversations: analysis of discourse in small laboratory groups. In: Schnase JL, Cunnius EL *Proceedings of CSCL ‘95: The First International Conference on Computer Support for Collaborative Learning*. New Jersey: Lawrence Erlbaum Associates Inc, pages 204–208.

Kelly GJ, Crawford T (1996) Students’ interaction with computer representations: analysis of discourse in laboratory groups. *Journal of Research in Science Teaching* **33:** 693–707.

*Kempa RF, Ayob A (1991) Learning interactions in group work in science. *International Journal of Science Education* **13:** 341–354.

Kempa RF, Ayob A (1995) Learning from group work in science. *International Journal of Science Education* **17:** 743–754.

*Keys CW (1995) An interpretive study of students' use of scientific reasoning during a collaborative report writing intervention in ninth grade general science. *Science Education* **79:** 415–435.

Keys CW (1997) An investigation of the relationship between scientific reasoning, conceptual knowledge and model formulation in a naturalistic setting. *International Journal of Science Education* **19:** 957–970.

Keys CW (1998) A study of grade six students generating questions and plans for open-ended science investigations. *Research in Science Education* **28:** 301–316.

Kneser C, Ploetzner R (2001) Collaboration on the basis of complementary domain knowledge: observed dialogue structures and their relation to learning success. *Learning and Instruction* **11:** 53–83.

Kortland K (1996) An STS case study about students' decision making on the waste issue. *Science Education* **80:** 673–689.

Kumpulainen K, Salovaara H, Mutanen M (2001) The nature of students' sociocognitive activity in handling and processing multimedia-based science material in a small group learning task. *Instructional Science* **29:** 481–515.

Kurth LA, Anderson CW, Palincsar AS (2002) The case of Carla: dilemmas of helping all students to understand science. *Science Education* **86:** 287–313.

Lajoie SP, Lavigne NC, Guerrera C, Munsie SD (2001) Constructing knowledge in the context of BioWorld. *Instructional Science* **29:** 155–186.

Lavoie DR (1999) Effects of emphasizing hypothetico-predictive reasoning within the science learning cycle on high school students' process skills and conceptual understandings in biology. *Journal of Research in Science Teaching* **36:** 1127–1147.

Lazarowitz R, Hertz RL, Baird JH, Bowlden V (1988) Academic achievement and on-task behavior of high school biology students instructed in a cooperative small investigative group. *Science and Education* **72:** 475–487.

Lonning RA (1993) Effect of cooperative learning strategies on student verbal interactions and achievement during conceptual change instruction in 10th grade general science. *Journal of Research in Science Teaching* **30:** 1087–1101.

Looi CK, Ang D (2000) A multimedia-enhanced collaborative learning environment. *Journal of Computer Assisted Learning* **16:** 2–13.

Lumpe AT, Staver JR (1995) Peer collaboration and concept development: learning about photosynthesis. *Journal of Research in Science Teaching* **32:** 71–98.

Matheson D, Achterberg C (2001) Ecologic study of children's use of a computer nutrition education program. *Journal of Nutrition Education* **33:** 2–9.

McKittrick B, Mulhall P, Gunstone R (1999) Improving understanding in physics: an effective teaching procedure. *Australian Science Teachers Journal* **45:** 27–33.

Meyer K, Woodruff E (1997) Consensually driven explanation in science teaching. *Science Education* **81:** 173–192.

Mortimer EF (1998) Multivoicedness and univocality in classroom discourse: an example from theory of matter. *International Journal of Science Education* **20:** 67–82.

Osborne J, Duschl RA, Fairbrother R (2002) *Breaking the Mould? Teaching Science for Public Understanding.* London: Nuffield Foundation.

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

Palincsar AS, Anderson C, David YM (1993) Pursuing scientific literacy in the middle grades through collaborative problem solving. *Elementary School Journal* **93:** 643–658.

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.

Pizzini EL, Shepardson DP (1992) A comparison of the classroom dynamics of a problem-solving and traditional laboratory model of instruction using path-analysis. *Journal of Research in Science Teaching* **29:** 243–258.

*Ploetzner R, Fehse E, Kneser C, Spada H (1999) Learning to relate qualitative and quantitative problem representations in a model-based setting for collaborative problem solving. *Journal of the Learning Sciences* **8:** 177–214.

Ratcliffe M (1997) Pupil decision-making about socio scientific-issues within the science curriculum. *International Journal of Science Education* **19:** 167–182.

Richmond G, Striley J (1996) Making meaning in classrooms: social processes in small-group discourse and scientific knowledge building. *Journal of Research in Science Teaching* **33:** 839–858.

Ritchie SM, Tobin K (2001) Actions and discourses for transformative understanding in a middle school science class. *International Journal of Science Education* **23:** 283–299.

Robblee KM (1991) Cooperative chemistry. Make a bid for student involvement. *Science Teacher* **58:** 20–23.

Roschelle J (1996) Learning by collaborating: Convergent conceptual change. In: Koschmann T (ed.) *CSCL: Theory and Practice of an Emerging Paradigm. Computers, Cognition and Work*. New Jersey: Lawrence Erlbaum Associates Inc.

*Roth WM (1994a) Science discourse through collaborative concept mapping: new perspectives for the teacher. *International Journal of Science Education* **16:** 437–455.

*Roth WM (1994b) Student views of collaborative concept mapping: an emancipatory research-project. *Science Education* **78:** 1–34.

*Roth WM (1996) The co-evolution of situated language and physics knowing. *Journal of Science Education and Technology* **5:** 171–191.

Roth WM (1999) Discourse and agency in school science laboratories. *Discourse Processes* **28:** 27–60.

Roth WM (2000) From gesture to scientific language. *Journal of Pragmatics* **32:** 1683–1714.

Roth WM, Duit R (2003) Emergence, flexibility and stabilization of language in a physics classroom. *Journal of Research in Science Teaching* **40:** 869–897.

Roth WM, McGinn MK, Woszczyna C, Boutonne S (1999) Differential participation during science conversations: the interaction of focal artifacts, social configurations, and physical arrangements. *Journal of the Learning Sciences* **8:** 293–347.

Roth WM, Roychoudhury A (1992) The social construction of scientific concepts or the concept map as conscription device and tool for social thinking in high school science. *Science and Education* **76:** 531–557.

Roth WM, Roychoudhury A (1993) The concept map as a tool for the collaborative construction of knowledge: a microanalysis of high-school physics students. *Journal of Research in Science Teaching* **30:** 503–534.

Roth WM, Welzel M (2001) From activity to gestures and scientific language. *Journal of Research in Science Teaching* **38:** 103–136.

Roth WM, Woszczyna C, Smith G (1996) Affordances and constraints of computers in science education. *Journal of Research in Science Teaching* **33:** 995–1017.

Roychoudhury A, Roth WM (1996) Interactions in an open-inquiry physics laboratory. *International Journal of Science Education* **18:** 423–445.

Russell DW, Lucas KB, McRobbie CJ (2003) The role of the microcomputer-based laboratory display in supporting the construction of new understandings in kinematics. *Research in Science Education* **33:** 217–243

Seiler G, Tobin K, Sokolic J (2001) Design, technology, and science: sites for learning, resistance and social reproduction in urban schools. *Journal of Research in Science Teaching* **38:** 746–767.

She HC (1999) Students' knowledge construction in small groups in the seventh grade biology laboratory: verbal communication and physical engagement. *International Journal of Science Education* **21:** 1051–1066.

Sherman GP, Klein JD (1995a) The effects of cued interaction and ability grouping during cooperative computer-based science instruction. *Educational Technology Research and Development* **43:** 5–24.

*Sherman GP, Klein JD (1995b) The effects of cued interaction and ability grouping during cooperative computer-based science instruction. Paper presented at the Annual Meeting of the American Educational Research Association. San Francisco, CA, USA: April 18–22.

Smeh K, Fawns R (2000) Classroom management of situated group learning: a research study of two teaching strategies. *Research in Science Education* **30:** 225–240.

*Solomon J (1991) Group discussions in the classroom. *School Science Review* **72:** 29–34.

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.

*Solomon J, Harrison K (1990) Arguing about industrial wastes. *Education in Chemistry* **27:** 160–162.

*Solomon J, Harrison K (1991) Talking about science based issues: do boys and girls differ? *British Educational Research Journal* **17:** 283–294.

Stein M (1997) Lightly stepping into science. *Science and Children* **34:** 18–21.

Suthers D, Weiner A (1995) Groupware for developing critical discussion skills. In: Schnase JL, Cunnius EL (eds) *Proceedings of CSCL '95: The First International Conference on Computer Support for Collaborative Learning*. New Jersey: Lawrence Erlbaum Associates Inc, pages 341–348.

Taconis R, Van Hout-Wolters B (1999) Systematic comparison of solved problems as a cooperative learning task. *Research in Science Education* **29:** 313–339.

Tao PK (1999) Peer collaboration in solving qualitative physics problems: the role of collaborative talk. *Research in Science Education* **29:** 365–383.

Tao PK (2000a) Computer supported collaborative physics learning: Developing understanding of image formation by lenses. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching. New Orleans, LA, USA: April 28–May 1.

*Tao PK (2000b) Developing understanding through confronting varying views: The case of solving qualitative physics problems. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching. New Orleans, LA, USA : April 28–May 1.

Tao PK (2001) Developing understanding through confronting varying views: the case of solving qualitative physics problems. *International Journal of Science Education* **23:** 1201–1218.

Tao PK (2003) Eliciting and developing junior secondary students' understanding of the nature of science through a peer collaboration instruction in science stories. *International Journal of Science Education* **25:** 147–171.

Tao PK, Gunstone RF (1999) Conceptual change in science through collaborative learning at the computer. *International Journal of Science Education* **21:** 39–57.

Teasley SD, Roschelle J (1993) Constructing a joint problem space: the computer as a tool for sharing knowledge. In: Lajoie SP, Derry SJ (eds) *Computers as Cognitive Tools. Technology in Education*. New Jersey: Lawrence Erlbaum Associates Inc.

Theberge CL (1994) Small-group vs. whole-class discussion: Gaining the floor in science lessons. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA, USA: April 7.

Tiberghien A, de Vries E (1997) Relating characteristics of teaching situations to learner activities. *Journal of Computer Assisted Learning* **13:** 163–174.

Tingle JB, Good R (1990) Effects of cooperative grouping on stoichiometric problem solving in high school chemistry. *Journal of Research in Science Teaching* **27:** 671–683.

Tolmie A, Howe C (1993) Gender and dialogue in secondary school physics. *Gender and Education* **5:** 191–209.

Tomkins SP, Dale S (2001) Looking for ideas: observation, interpretation and hypothesis-making by 12-year-old pupils undertaking science investigations. *International Journal of Science Education* **23:** 791–813.

Tsai CC (1999) 'Laboratory exercises help me memorize the scientific truths': a study of eighth graders' scientific epistemological views and learning in laboratory activities. *Science and Education* **83:** 654–674.

Van Boxtel C, Van der Linden J, Kanselaar G (2000a) Collaborative learning tasks and the elaboration of conceptual knowledge. *Learning and Instruction* **10:** 311–330.

Van Boxtel C, Van der Linden J, Kanselaar G (2000b) The use of textbooks as a tool during collaborative physics learning. *Journal of Experimental Education* **69:** 57–76.

*Van Boxtel C, Roelofs E (2001) Investigating the quality of student discourse: what constitutes a productive student discourse? *Journal of Classroom Interaction* **36:** 55–62.

*Van Boxtel C, Van der Linden J, Kanselaar G (1997) Collaborative construction of conceptual understanding: interaction processes and learning outcomes emerging from a concept mapping and a poster task. *Journal of Interactive Learning Research* **8:** 341–361.

Van Zee EH, Iwasyk M, Kurose A, Simpson D, Wild J (2001) Student and teacher questioning during conversations about science. *Journal of Research in Science Teaching* **38:** 159–190.

Vellom RP, Anderson CW, Palincsar AS (1995) Developing mass, volume and density as mediational means in a sixth grade classroom. Paper presented at the Annual Meeting of the American Educational Research Association. San Francisco, CA, USA: April 18–22.

*Vellom RP, Anderson CW, Palincsar AS (1994) Constructing facts and mediational means in a middle school science classroom. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA, USA: May 24.

Webb NM, Nemer KM, Chizhik AW, Sugrue B (1998) Equity issues in collaborative group assessment: group composition and performance. *American Educational Research Journal* **35:** 607–661.

*Webb NM, Nemer KM, Chizhik AW, Sugrue B (1995) *Using Group Collaboration as a Window into Students' Cognitive Processes*. Los Angeles, CA, USA: National Center for Research on Evaluation, Standards and Student Testing.

*Webb NM, Nemer KM, Zuniga S (2002) Short circuits or superconductors? Effects of group composition on high-achieving students' science assessment performance. *American Educational Research Journal* **39:** 943–989.

Wellington J, Osborne J (2001) Discussion in school science: learning science through talking. In: Wellington J, Osborne J (eds) *Language and Literacy in Science Education*. Milton Keynes: Open University Press, pages 82–102.

Whitelock D, Scanlon E, Taylor J, O'Shea T (1995) Computer support for pupils collaborating: a case study on collisions. In: Schnase JL, Cunnius EL (eds) *Proceedings of CSCL '95: The First International Conference on Computer Support for Collaborative Learning*. New Jersey: Lawrence Erlbaum Associates Inc, pages 380–384.

Williams A (1995) Long-distance collaboration: a case study of science teaching and learning. In: Spiegel SA (ed.) *Perspectives from Teachers' Classrooms. Action Research. Science FEAT (Science for Early Adolescence Teachers).* Tallahassee, FL, USA: Southeastern Regional Vision for Education.

Windschitl M (2001) Using simulations in the middle school: does assertiveness of dyad partners influence conceptual change? *International Journal of Science Education* **23:** 17–32.

Woodruff E, Meyer K (1997) Explanations from intra- and inter-group discourse: students building knowledge in the science classroom. *Research in Science Education* **27:** 25–39.

Zohar A, Nemet F (2002) Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching* **39:** 35–62.

6.2 Other references used in the text of the report

Aronson E, Stephen C, Sikes J, Blaney N, Snapp M (1978) *The Jigsaw Classroom*. California: Sage.

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

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.

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

Bentley D, Watts M (1989) *Learning and teaching in school science: practical alternatives.* Buckingham: Open University Press.

Campbell B, Kaunda L, Allie S, Buffler A, Lubben F (2000) The communication of laboratory investigations by university entrants. *Journal of Research in Science Teaching* **37:** 839–853.

Daws N, Singh B (1999) Formative assessment strategies in secondary science. *School Science Review* **80:** 71–78.

Department for Education and Employment (DfEE) (1998) *The National Literacy Strategy*. London: DfEE.

Department for Education and Science (DfES) (1999) *Science: The National Curriculum for England*. London: DfES/Qualifications and Curriculum Authority (QCA).

Driver R, Guesne E, Tiberghien A (eds) (1985) *Children's ideas in science*. Buckingham: Open University Press.

Driver R, Asoko, H, Leach J, Mortimer E, Scott P (1994) Constructing scientific knowledge in the classroom. *Educational Researcher* **23:** 5–12.

EPPI-Centre (2002a) *EPPI-Centre Core Keywording Sheet (Version 0.9.7*). London: EPPI-Centre, Social Science Research Unit.

EPPI-Centre (2002b*) EPPI-Centre Core Keywording Strategy (Version 0.9.7).* London: EPPI-Centre, Social Science Research Unit.

EPPI-Centre (2002c) *EPPI-Centre EPPI-Reviewer (Version 0.9.7)*. London: EPPI-Centre, Social Science Research Unit.

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

Gott R, Duggan S (1996) Practical work: its role in the understanding of evidence in science. *International Journal of Science Education* **18:** 791–806.

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.

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

Levinson R, Turner S (2001) *Valuable Lessons: Engaging with the Social Context of Science in Schools*. London: The Wellcome Trust.

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

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.

Osborne J, Duschl R, Fairbrother R (2002) *Breaking the Mould? Teaching Science for Public Understanding*. London: The Nuffield Foundation.

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

Solomon J, Scott L, Duveen J (1996) Large-scale exploration of pupils' understanding of the nature of science. *Science Education* **80:** 493–508.

APPENDIX 1.1: Consultancy Group membership

The Review Group for Science benefits from the advice of a group of national and international consultants, all with expertise in particular areas and aspects of science education.

- Professor Nancy Brickhouse, University of Delaware, USA, and editor of *Science Education*
- Professor Rick Duschl, King's College, University of London, UK, and former editor of *Science Education*
- Mike Driver, Inspector at the Office for Standards in Education (Ofsted) and Science Inspector for Cleveland Local Education Authority, UK
- Chris Edwards, Chief Education Officer, Leeds, UK
- Josette Farrugia, University of Malta and Schools Examinations Officer for Science
- Peter Finegold, Education Office for the Wellcome Trust
- Professor John Gilbert, University of Reading, UK, and editor of the *International Journal of Science Education*
- Professor John Leach, University of Leeds, UK
- Peter Nicolson, University of York Science Education Group, UK
- Colin Osborne, Education Officer, Royal Society of Chemistry, UK
- Professor Jonathan Osborne, King's College, University of London, UK
- Professor Manfred Prenzel, Leibniz Institute for Science Education (IPN), University of Kiel, Germany
- Professor Michael Reiss, Institute of Education, University of London, UK
- Professor Marissa Rollnick, University of the Witwatersrand, Johannesburg, South Africa
- Miranda Stephenson, Chemical Industries Education Centre, University of York, UK
- Nigel Thomas, Education Officer at the Royal Society, UK

APPENDIX 2.1: Inclusion and exclusion criteria

Inclusion and exclusion criteria were applied hierarchically.

Systematic review question:
How are small-group discussions used in science teaching with students aged 11-18, and what are the effects on students' understanding in science or attitudes to science?

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. NOT mathematics, technology, social science or computing*

2. **Not about the use of group discussions**
   – *Includes both synchronous and asynchronous group discussion (e.g. computer- mediated)*

3. **Not about small-groups**
   – *Two to six participants*

4. **Not on substantive and explicit discussion tasks**
   – *Explicit discussion tasks taking more than two minutes*

5. **If only about effects of group discussions, *not* about the effect on students' understanding or attitude**
   – *'Understanding' includes understanding of science concepts and ideas about science*
   – *'Attitude' includes attitude to science and to science education*

6. **Not about learners aged 11 to 18, or main focus not on learners aged 11 to 18**
   – *Out of school can be included*

EXCLUSION ON STUDY TYPE

7. (a) Editorials, commentaries, book reviews or position papers
   (b) Policy documents, syllabuses, frameworks or specifications
   (c) Resources
   (d) Bibliography
   (e) Theoretical (non-empirical) paper
   (f) Methodology paper

EXCLUSION ON SETTING IN WHICH STUDY WAS CARRIED OUT

8. Not published in English

9. Not published in the period 1980–2002

APPENDIX 2.2: Search strategy for electronic databases

Subject
Small-group discussions in science teaching

Population
Students aged 11 to 18

Limits
English language
1980 to 2003

2.2.1 Educational Resources Information Center (ERIC)

ERIC was searched on 27 February 2003, using the BIDS Ovid interface and 836 records were retrieved.

1 exp cooperative learning/
2 "ARGUMENTATION".mp.
3 exp discourse analysis/ or exp persuasive discourse/
4 exp discussion/ or exp "discussion (teaching technique)"/ or exp discussion groups/ or exp group discussion/
5 1 or 2 or 3 or 4
6 5 and (science or biology or chemistry or physics or earth science).mp. [mp=abstract, title, headings word, identifiers, full text]
7 limit 6 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 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

The search was updated on 4 May 2004, this time using the Dialog interface but working with the available subject headings and related terms in order to match the original search as closely as possible. A further 148 records were retrieved.

1 'cooperative learning' or 'small group instruction' or 'learning strategies' or 'group discussion'
2 'argumentation' or 'verbal communication' or 'discourse analysis'
3 'persuasive discourse' or 'persuasive strategies'
4 'discussion groups' or 'discussion (teaching technique)' or 'discussion'
5 1 or 2 or 3 or 4
6 5 and ('science' or 'biology' or 'chemistry' or 'physics' or 'earth science')
7 limit 6 to ('english language') and ('secondary education' or 'elementary education' or 'intermediate grades' 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 – proceedings' or 'dissertations/theses' or 'reports – research' or 'journal articles' or 'speeches/meeting papers') and yr=2003

2.2.2 British Education Index (BEI)

BEI was searched on 27 February 2003, using the BIDS Ovid interface and 56 records were retrieved.

1 cooperative learning.mp. [mp=title, edition statement, abstract, heading word]
2 argumentation.mp. [mp=title, edition statement, abstract, heading word]
3 exp discourse analysis/ or exp persuasive discourse/
4 exp discussion/ or exp "discussion (teaching technique)"/ or exp discussion groups/ or exp group discussion/
5 1 or 2 or 3 or 4
6 exp group dynamics/ or exp group work/ or exp small group teaching/ or "group dynamics or small group teaching".mp.
7 5 or 6
8 7 and (science or biology or chemistry or physics or earth science).mp. [mp=title, edition statement, abstract, heading word]
9 limit 8 to (english and (primary secondary education or middle school education or secondary education or sixth form education or sixteen to nineteen education or further education))

The search was updated on 4 May 2004 using the Dialog interface. The original search was matched but was also enhanced by the exploration of additional subject headings and related terms. A further 27 records were retrieved.

1 'cooperative learning' or 'small group instruction' or 'learning strategies' or 'group discussion'
2 'argumentation' or 'verbal communication' or 'discourse analysis'
3 'persuasive discourse' or 'persuasive strategies'
4 'discussion groups' or 'discussion (teaching technique)' or 'discussion'
5 1 or 2 or 3 or 4
6 5 and ('science' or 'biology' or 'chemistry' or 'physics' or 'earth science')
7 limit 6 to ('english language') and ('secondary education' or 'elementary education' or 'intermediate grades' or 'middle schools' or 'junior high schools' or 'high schools' or 'high school equivalency programs' or 'postsecondary education' or 'two year colleges') and yr=2003

2.2.3 PsycINFO

PsycINFO was searched on 10 April 2003, using the WEBSPIRS interface and 537 records were retrieved. For the reasons given in section 2.2.2, this search was not updated in 2004.

1 (cooperative-learning or cooperation or cooperation- or cooperative) in MJ,MN,AG,PO,KC
2 (argument or argumentation) in MJ,MN,AG,PO,KC
3 (discourse-analysis or discourse-processes or discourses) in MJ,MN,AG,PO,KC

4 (discussion-group or group-decision-making or group-discussion or group-dynamics or group-decision-and-negotiation) in MJ,MN,AG,PO,KC
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 Limit 6 to (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)

2.2.4 Social Science Citation Index (SSCI)

SSCI was searched on 16 April 2003, using the Web of Science interface and 568 records were retrieved. The search was updated using the same interface on 4th May 2004 and a further 74 records were retrieved.

1 (cooperative or collaborative) and (science or biology or chemistry or physics or earth science) and (student* or pupil* or learner*)
2 (argumentation or discourse) and (science or biology or chemistry or physics or earth science) and (student* or pupil* or learner*)
3 (small group*) and (science or biology or chemistry or physics or earth science) and (student* or pupil* or learner*)
4 1 or 2 or 3
5 Limit 4 to English and articles

APPENDIX 2.3: Journals handsearched

The following key journals were handsearched for potentially relevant papers:

Journal of Biological Education

Journal of Chemical Education

Research in Science and Technological Education

Research in Science Education

Studies in Science Education

Other key journals were found to be indexed to one or more of the electronic databases and were therefore fully covered by the electronic searches. These were as follows:

British Journal of Developmental Psychology

Cognition and Instruction

Discourse Processes

Instructional Science

International Journal of Science Education (formerly the *European Journal of Science Education*)

Journal of Educational Research

Journal of Research in Science Teaching

Learning and Instruction

Physics Education

School Science Review

Science Education

APPENDIX 2.4: EPPI-Centre keyword sheet including review-specific keywords

EPPI-CENTRE EDUCATIONAL KEYWORDING SHEET V0.9.7 *Bibliographic details and/or unique identifier*……………………………

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

2. Status
Published
In press
Unpublished

3. 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.)
……………………………………………………
……………………………………………………
……………………………………………
…………………………………………………

4. **Language** (Please specify.)

……………………………………………

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

………………………………………………

6. 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.)

7. 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.)………………………..

8. Programme name (Please specify.)
……………………………………………

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

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

11. Sex of learners
Female only
Male only
Mixed sex

12. 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………………………

13. Which type(s) of study does this report describe?
a. Description
b. Exploration of relationships
c. Evaluation
- naturally-occurring
- researcher-manipulated*

d. Development of methodology
e. Review
f. Systematic review
g. Other review
**see 14.*

14. To assist with the development of a trials register please state if a researcher-manipulated evaluation is one of the following:

Controlled trial (non-randomised)
Randomised controlled trial (RCT)

Please state here if keywords have not been applied from any particular category and the reason why (e.g. no information provided in the text).

……………………………………………
……………………………………………
……………………………………………
……………………………………………
……………………………………………
………………………………………

Keyworded by……………………………………..

Date…………………………………

Review-specific keywords *For each item tick any number of keywords*

15. Does the study focus on the effects of small-group discussions?

a. *No*, but on the *use* of small-group discussions
b. *Yes*, on the *effect on understanding* of science
c. *Yes*, on the *effect on attitudes* to science

16. What discipline?

a. (integrated) Science
b. Biology
c. Chemistry
d. Physics
e. Earth science

17. What types of learners are involved?

a. mixed ability
b. lower ability / slow learners
c. upper ability / gifted
d. disaffected
e. unspecified
f. other: ………………………………………….

18. What is the mode of group discussions?

a. synchronous (i.e. face-to-face)
b. asynchronous (i.e. IT-mediated)

19. How are discussion groups constituted?

a. friendship ties, i.e. learners' choice
b. randomly, by teacher
c. randomly, but same sex groups
d. purposely same ability
e. purposely heterogeneously
f. other: ………………………………………….

20. What is the size of the discussion groups?

a. 2 (dyads)
b. 3 or 4
c. 5 or 6
d. unspecified

21. What is the stimulus for discussion tasks?

a. one-line oral teacher instruction
b. oral context provided by teacher only
c. newspaper article
d. prepared curriculum print materials
e. practical work
f. computer software
g. field trip
h. video/TV/film clip
i. learner-generated
j. other: ………………………………………….

22. What is the duration of discussion tasks?

a. 2-5 minutes
b. 6-30 minutes
c. close to a class period (30-60 minutes)
d. longer than a class period
e. unspecified

23. What is the organisation of discussion tasks?

a. self-contained
b. accretion (snowballing) 2 > 4 > 8
c. jigsawing
d. envoying
e. other: ………………………………………….

24. What is the product of the discussion tasks?

a. individual sense-making
b. report group views/presentation orally in class
c. support a group position in a class debate / quiz
d. present group written project (incl. poster)
e. other: ………………………………………….

25. How many discussion groups are included?

a. 1 discussion group only
b. 2 discussion groups
c. 3–10 discussion groups
d. 11–30 discussion groups
e. more than 30 discussion groups
f. unspecified

26. Outcomes are reported in terms of:

a. conceptual understanding of science
b. evidence (methods and nature of science)
c. applications of science
d. attitudes to (school) science
e. skills (communication/collaboration)
f. decision-making on socio-scientific issues

For learners of different:

g. ability (lower / middle / higher)
h. gender
i. educational level

27. What is the research strategy:

a. experiment
b. survey
c. case study
d. action research
e. ethnography

28. What is the nature of the data?

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 (audio)
h. presentations
i. observed behaviour (including video)
j. computer logs

APPENDIX 2.2.5: Weight of evidence indicators

Review question: When using different stimuli (print materials, practical work, ICT, video/film), what is the evidence from evaluative studies of the effect of small-group discussions (SGD) on students' understanding of evidence in science?

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 M11, B and C: what is the overall weight of evidence this study provides to answer *this review question*?
high (3)	medium (2)	low (1)	high (3)	medium (2)	low (1)	If equal weighting of M11, B and C, each weighted across the range as low (1), medium low (2), medium (3), medium high (4) and high (5)
For the RQs relevant to the review:			***For the RQs relevant to the review:***			
Sample size			**Nature of sample**			**Sum total and classification for D** 3–4: low 5–7: medium low 8–10: medium 11–13: medium high 14–15: high
large sample with appr. sampling method	large sample, no sampling method	small sample (up to three classes)	highly representative of the type of stimulus	less representative of the type of stimulus	not representative of the type of stimulus	
Comparison/ control			**Focus of intervention**			
comparison for stimuli in SGD in design (control, types)	comparison for stimuli in SGD in findings only	no comparison/control	Stimulus is the main independent variable.	Stimulus is a major discrete element of intervention.	Stimulus is wrapped up in intervention.	
Benchmark data			**Measures**			
pre-post data on understanding of evidence	longitudinal dev of understanding of evidence	only post data for understanding of evidence	highly appropriate for testing understanding of evidence directly	mildly appropriate for testing understanding of evidence directly	appropriate for testing understanding of evidence indirectly	
Data collection			**Breadth**			
solid checks on reliability/validity for data collection	some checks on reliability/validity for data collection	few/no checks on reliability/validity for data collection	reports broad range of understanding of evidence	reports narrow range of understanding of evidence	reports understanding of evidence only indirectly	
Data analysis			**Situation**			
solid checks on reliability/validity for data analysis	some checks on reliability/validity for data collection	few/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 3.1: Types of study included in the systematic map

Tables A–D tabulate all 94 studies in the review according to the type of research study reported.

Table A lists the 14 reports of descriptive studies.

Table B provides an overview of the 32 studies reporting explorations of relationships.

Tables C and D list the reports of the 23 naturally-occurring and 25 researcher-manipulated evaluation studies, respectively.

In line with the three aspects of the review question, for each paper the foci of the study are indicated: that is, the use of small-group discussions, the effect on understanding of science and the effect on attitudes to science. Equally, the tables specify the terms in which the findings are reported.

As stated before, the area of ‘understanding of science’ is divided into three sub-areas: that is, the understanding of science concepts, the understanding of evidence in science and the ability to apply science concepts. In addition, information is listed on reports of attitudinal aspects, communication skills of group members and decision-making skills on socio-scientific issues.

Table A: Summary of reports of descriptive studies included in the review (N = 14)

		Focus of study			**Findings reported in terms of**					
Record number	***Author and year***	***Use of small-group discussions***	***Effect on understanding***	***Effect on attitudes***	***Concepts***	***Evidence***	***Applications***	***Attitudes***	***Skills***	***Decision-making***
1067	McKittrick *et al.*, 1999	✓			✓				✓	
1334	Ritchie and Tobin, 2001	✓			✓				✓	
1378	Roth, 2000	✓			✓				✓	
1384	Roth and Roychoudhury, 1993	✓			✓				✓	
1823	Wellington and Osborne, 2001	✓			✓					
1322	Richmond and Striley, 1996	✓				✓				
481	Fawns and Salder, 1996	✓							✓	
977	Looi and Ang, 2000	✓							✓	
1377	Roth, 1999	✓							✓	
1398	Roychoudhury and Roth, 1996	✓							✓	
570	Goldman, 1996	✓	✓		✓				✓	
1355	Roschelle, 1996		✓		✓				✓	
2045	Roth and Duit, 2003		✓		✓					
1183	Osborne *et al.*, 2001		✓			✓				✓

Table B: Summary of reports of studies exploring relationships included in the review (N = 32)

		Focus of study			**Findings reported in terms of**					
Record number	***Author and year***	***Use of small-group discussions***	***Effect on understanding***	***Effect on attitudes***	***Concepts***	***Evidence***	***Applications***	***Attitudes***	***Skills***	***Decision-making***
900	Kurth *et al.*, 2002	✓			✓	✓			✓	
1033	Matheson and Achterberg, 2001	✓			✓					
1597	Theberge, 1994	✓			✓				✓	
1607	Tiberghien and de Vries, 1997	✓			✓				✓	
1658	Van Zee *et al.*, 2001	✓			✓					
769	Jimenez *et al.*, 1998	✓				✓				

		Focus of study			**Findings reported in terms of**					
Record number	***Author and year***	***Use of small-group discussions***	***Effect on understanding***	***Effect on attitudes***	***Concepts***	***Evidence***	***Applications***	***Attitudes***	***Skills***	***Decision-making***
770	Jimenez *et al.*, 2000a	✓				✓				✓
779	Johnson and Stewart, 2002	✓				✓				
823	Kelly and Crawford, 1996	✓						✓		
1862	Keys, 1998	✓							✓	
502	Ford, 1999	✓							✓	
1387	Roth, 1996	✓							✓	
695	Hogan, 2002	✓	✓		✓					✓
1103	Mortimer, 1998	✓	✓		✓					
1382	Roth *et al.*, 1999	✓	✓		✓				✓	
1386	Roth and Welzel, 2001	✓	✓		✓				✓	
1584	Tao, 2000a	✓	✓		✓				✓	
1587	Tao and Gunstone, 1999	✓	✓		✓				✓	
1592	Teasley and Rochelle, 1993	✓	✓		✓				✓	
1622	Tomkins and Dale, 2001	✓	✓		✓					
767	Jimenez, 2002	✓	✓		✓	✓			✓	✓
1081	Meyer and Woodruff, 1997	✓	✓		✓	✓			✓	
1777	Woodruff and Meyer, 1997	✓	✓		✓	✓				
1678	Vellom *et al.*, 1995	✓	✓		✓	✓			✓	
1389	Roth and Roychoudhury, 1992	✓	✓		✓	✓			✓	
693	Hogan, 1999a	✓	✓		✓	✓		✓	✓	
1632	Tsai, 1999	✓		✓		✓		✓	✓	
1824	Osborne *et al.*, 2002	✓	✓	✓	✓	✓	✓	✓		✓
1457	Seiler *et al.*, 2001	✓	✓	✓	✓			✓	✓	
1514	Solomon, 1992	✓	✓	✓	✓			✓	✓	✓
2049	Russell *et al.*, 2003		✓		✓	✓				
1544	Stein, 1997		✓			✓				

Table C: Summary of reports of naturally-occurring evaluative studies included in the review (N = 23)

		Focus of study			**Findings reported in terms of**					
Record number	***Author and year***	***Use of small-group discussions***	***Effect on understanding***	***Effect on attitudes***	***Concepts***	***Evidence***	***Applications***	***Attitudes***	***Skills***	***Decision-making***
1	Hornsey and Horsfield, 1982	✓							✓	
539	Gayford, 1993	✓							✓	
553	Gilbert and Pope, 1986	✓							✓	
1821	Ratcliffe, 1997	✓								✓
39	Alexopoulou and Driver, 1996	✓	✓		✓				✓	
62	Arvaja *et al.*, 2000	✓	✓		✓					
781	Johnston and Scott, 1991	✓	✓		✓					
828	Kempa and Ayob, 1995	✓	✓		✓				✓	
883	Kortland, 1996	✓	✓		✓				✓	✓
993	Lumpe and Staver, 1995	✓	✓		✓				✓	
1610	Tingle and Good, 1990	✓	✓		✓				✓	
1582	Tao, 1999	✓	✓		✓					
1585	Tao, 2001	✓	✓		✓	✓			✓	
1197	Palincsar *et al.*, 1993	✓	✓		✓	✓				
374	De Vries *et al.*, 2002	✓	✓		✓	✓			✓	
842	Keys, 1997	✓	✓		✓	✓			✓	
492	Finkel, 1996	✓	✓		✓	✓				
1835	Suthers and Weiner, 1995	✓	✓			✓			✓	
133	Bianchini, 1997	✓	✓	✓	✓			✓	✓	
930	Lazarowitz *et al.*, 1988		✓		✓				✓	
1338	Robblee, 1991		✓	✓	✓		✓	✓		
1586	Tao, 2003		✓			✓				
1857	Williams, 1995		✓	✓	✓	✓		✓	✓	

Table D: Summary of reports of researcher-manipulated evaluative studies included in the review (N = 25)

		Focus of study			**Findings reported in terms of**					
Record number	***Author and year***	***Use of small-group discussions***	***Effect on understanding***	***Effect on attitudes***	***Concepts***	***Evidence***	***Applications***	***Attitudes***	***Skills***	***Decision-making***
741	Hynd *et al.*, 1994	✓	✓		✓					
868	Kneser and Ploetzner, 2001	✓	✓		✓				✓	
898	Kumpulainen *et al.*, 2001	✓	✓		✓					
1723	Webb *et al.*, 1998	✓	✓		✓				✓	
916	Lajoie *et al.*, 2001	✓	✓		✓	✓				
1619	Tolmie and Howe, 1993	✓	✓		✓	✓			✓	
1816	Zohar and Nemet, 2002	✓	✓		✓	✓				✓
1578	Taconis and Van Hout-Wolters, 1999		✓		✓				✓	
1836	Whitelock *et al.*, 1995		✓		✓					
254	Chang and Mao, 1999b		✓		✓					
253	Chang and Mao, 1999a		✓	✓	✓			✓		
541	Gayford, 1995		✓	✓	✓	✓	✓	✓		✓
926	Lavoie, 1999		✓	✓	✓	✓		✓		
Randomised controlled trials (N = 12)										
1243	Pizzini and Shepardson, 1992	✓							✓	
1467	She, 1999	✓							✓	
1861	Smeh and Fawns, 2000	✓							✓	
250	Chan, 2001	✓	✓		✓					
976	Lonning, 1993	✓	✓		✓				✓	
1649	Van Boxtel *et al.*, 2000b	✓	✓		✓				✓	
1648	Van Boxtel *et al.*, 2000a	✓	✓		✓				✓	
1761	Windschitl, 2001	✓	✓		✓				✓	
1218	Pederson, 1992	✓	✓	✓	✓			✓		
692	Hogan, 1999b	✓	✓	✓	✓	✓		✓		
1471	Sherman and Klein, 1995a		✓	✓		✓		✓	✓	
258	Chang and Lederman, 1994		✓		✓				✓	

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

1. Keys CW (1997) An investigation of the relationship between scientific reasoning, conceptual knowledge and model formulation in a naturalistic setting. *International Journal of Science Education* **19:** 957-970. 2. Keys CW (1995) An interpretive study of students' use of scientific reasoning during a collaborative report writing intervention in ninth grade general science. *Science Education* **79:** 415-435.	
Country of study	USA
Details of researchers	Doing a PhD at Georgia State University. A teacher and a university preservice intern also facilitated student work.
Name of programme	Not applicable
Age of learners	14 to 15
Type of study	Evaluation: naturally-occurring
Aims of study	To investigate the use of reasoning strategies through a collaborative writing task in order to generate meaningful scientific models, and the evidence for improvement in students' reasoning discourse
Summary of study design, including details of sample	• Pre- and post-intervention clinical interviews with four individual students regarding conceptual knowledge • Two single-sex pairs underwent the intervention and generated collaboratively a report for two laboratory activities. The domain-specific knowledge for one activity was low; for the other, high. • Reasoning strategies in interactions between pairs were video-recorded, and in individual and joint written products were collected. • The types of reasoning strategies resulting in conceptual change were identified. • For paper 2, no interviews were used and three pairs were involved. The types of reasoning strategies used were classified and their development over a three-month period traced. • Actual sample: paper 1– two pairs, four students; paper 2 – three pairs, six students
Methods used to collect data	• One-to-one interview: pre- and post-intervention clinical interviews • Observation: video-recorded pair interactions (two cameras!) • Self-completion questionnaire: written collaborative report of laboratory activity; written individual prior knowledge and predictions. • School/college records • Other documentation: researcher's field notes
Data-collection instruments, including details of checks on reliability and validity	• Sample of a reporting guideline is appended to paper 2. • No interview schedule is provided, but relevant interview responses are reported verbatim. • Checks on reliability: Triangulation of data sources (field notes, video footage, written records) increases reliability. • Checks on validity: This is an interpretive study, so the emphasis is on contextual validity. Extensive details are provided of the type of characteristics of students and the process of their involvement, the teaching procedures, and the context of the specific task being focused on. Some more detail on the general environment in the school would have been useful. • One task was used for development of a pilot collaborative report.
Methods used to analyse data, including details of checks on reliability and validity	• This is an interpretive study. Descriptive analysis: The domain-specific understanding in pre- and post-intervention interviews has been described according to the nature of concepts; accepted major types of misconceptions are used as classification. A constant comparative method is used for analysing the student interactions and written work for identifying similar reasoning strategies (paper 2, p 421) and patterns of scientific reasoning. For this, Kuhn's framework has been used and extended.

	• Assertions were created based on patterns in the data. • Checks on reliability: Independent coding of reasoning strategies of 13 units (10%) by two researchers with initial inter-coder agreement of 85%, and additional 11% no discussion. • Checks on validity: triangulation of three sources of data • Use of Kuhn's framework as starting point for analysis for strategies
Summary of results	*Paper 1* • RQ1: Across laboratory activities, the following types of reasoning were used: (a) recognising that prior ideas (models) may be incorrect; (b) evaluating new observations for consistency with current ideas and using evidence to modify ideas; and (c) coordinating all mutually consistent knowledge propositions into a coherent model. • RQ2: A comparison is not really made between the reasoning strategies employed in activities with low and high domain-specific demands respectively. However, the reasoning strategies used for each of these activities have been listed and illustrated. *Paper 2* • RQ3: Scientific reasoning can be identified by 11 skills clustered in four categories of reasoning skills for: (a) assessing prior models (posing predictions; evaluating predictions; explaining/justifying predictions); (b) generating new models (evaluating observations; identifying patterns; drawing conclusions; formulating models); (c) extending models (inferring; comparing/contrasting); and (d). supporting (discussing concept meaning; identifying relevant information). RQ4: The greatest improvement in reasoning discourse occurs in pairs who are initially reluctant to discuss the meaning of scientific concepts.
Conclusions	Teaching implications are discussed. The relationship of the findings to Kuhn's model is discussed.
Weight of evidence A: (trustworthiness in relation to study questions)	Medium-high Within the limitations set by the author (no generalisibility, interpretive design), the findings have a high-medium trustworthiness.
Weight of evidence B: Stimuli (appropriateness of research design and analysis)	Medium-low Very small sample size; no comparison/control; the benchmark data related to conceptual understanding rather than the use of evidence; data collection had no reliability checks, but context validity of this interpretive study is high; some methods were used to check data analysis.
Weight of evidence C: Stimuli (relevance of focus of study to review)	Medium Use of report prompts as stimulus is quite a specific way of using printed materials as stimulus; stimulus is not a tested independent variable; measures of change (pre-post tests) focus on conceptual knowledge change, rather than understanding evidence; the understanding of evidence is synthesised as the ability to build models; setting is very naturalistic.
Weight of evidence D: Stimuli (overall weight of evidence)	Medium

Lajoie SP, Lavigne NC, Guerrera C, Munsie SD (2001) Constructing knowledge in the context of BioWorld. *Instructional Science* **29:** 155-186.	
Country of study	Assumed Canada
Details of researchers	Researchers at McGill University, Canada funded by Canadian Sciences and Humanities Research Council and Wisconsin Alumni Research Fund
Name of programme	BioWorld computer program/software
Age of learners	14 to15
Type of study	Evaluation: researcher-manipulated
Aims of study	To examine students' use of Bioworld Computer learning environment to solve problems related to the digestive system and analyse how the student actions and verbal dialogue were conducted to pinpoint the types of features within BioWorld that were most conducive to learning and scientific reasoning
Summary of study design, including details of sample	Students from two grade 9 biology classes worked in pairs to use the BioWorld program. Classes were of comparable ability level. They were allowed to choose their own partners for the task. The entire sample was used for the first two research questions. Data from six pairs were used for research question 3 (role of teacher guided groups and of researcher guided group). Teacher selected these groups as being equivalent in terms of their previous grades and ability to articulate their understanding. Actual sample: 40 students
Methods used to collect data	• Observation: audio and video tapes • Computer log of actions and decisions on the BioWorld program
Data-collection instruments, including details of checks on reliability and validity	• Limited details are given; data about the students' choices about the diagnosis and how these changed, about access to virtual tests and other information was collected via the computer software. • Checks on reliability: Not explicitly stated but computer records and audio/video recordings are reliable and standard tools for this kind of research. • Checks on validity: Data from medical experts and teachers (not the teacher used in the intervention addressing RQ3) were used as benchmarks for indicators of student performance in scientific reasoning.
Methods used to analyse data, including details of checks on reliability and validity	Verbal data was not analysed but used as exemplars to support computer data. Statistical for computer data. • Initial one-way MANOVA test was used to determine if there was a difference between students from the two different classes. • A Pearson correlation was used for the features in terms of the relationship between group and expert actions. • A MANOVA test was used to investigate the condition (3) effects of instruction on all dependent measures of interest. • Checks on reliability: Included (i) statistical compensation for small sample size; (ii) statistical test to check to see if class variable is present; and (iii) a qualitative analysis of the verbal data from the two coached conditions demonstrated that a cognitive apprenticeship approach (Collins, Brown and Newman, 1988) to instruction was used by both teacher and graduate student. • Checks on validity: Not explicitly stated but used appropriate test for the data
Summary of results	Research question (RQ) 1: Groups versus expert use of BioWorld features • There was a significant correlation between proportion of expert symptoms collected during problem representation and overall evidence collected that was expert-like (r = 0.59, p = 0.002). • Declarative knowledge acquired was positively correlated with the proportion of expert-like diagnostic tests ordered (r = 0.42, p = 0.04). Hence declarative and procedural knowledge as defined in this study were correlated. • Those who scored highly on collecting expert evidence also scored highly on expert-like diagnostic tests ordered (r = 49, p = 0.02). RQ 2: Relationship between confidence and argumentation and diagnostic accuracy

	• Students significantly increased their confidence about their diagnosis at the time of their final argument. This was tied to final diagnostic accuracy but not to first hypothesis. As accuracy increased, confidence increased. RQ 3: Exploration of coaching styles and lack of coach. Only six pairs were used; qualitative analysis. • Teacher and graduate student used cognitive apprenticeship approach with some small differences in the amount of direction given depending on the particular student pairs. • Students working on BioWorld without adult support spent more time at the beginning on insignificant details but benefited from generating their own hypotheses and followed up on their own problem-solving strategies.
Conclusions	RQ 1 • BioWorld teaches students about the processes of scientific reasoning and demonstrates that students can learn about diseases efficiently. • Students who learned to reason scientifically took less time and needed fewer actions than students who did not make accurate diagnosis, indicating that the types of search strategies used by successful students were different from those used by less successful students. • The argumentation and reasoning patterns collected with BioWorld support the research on collaborative learning in that sophisticated patterns of scientific reasoning were found in small-group learning situations. RQ 2 • A strong relationship between student confidence and knowledge was found. As students acquired knowledge, their diagnoses increased dynamically within the environment. Confidence is a true indicator of students' diagnostic accuracy. RQ 3 • There were some differences in tutoring strategies between a teacher and a graduate student.
Weight of evidence A: (trustworthiness in relation to study questions)	Medium Medium-high for quantitative aspects; medium-low for qualitative aspects
Weight of evidence B: Stimuli (appropriateness of research design and analysis)	Low Small sample size and no sampling method; no control/comparison for stimulus, but for scaffolding by teacher; no pre-post testing; no information concerning reliability and validity of data collection; some validity check for data analysis; ANOVA data for conceptual understanding only; triangulation and the use of grounded theory for generating categories
Weight of evidence C: Stimuli (relevance of focus of study to review)	Medium-high Stimulus (ICT software with cues) representative for this stimulus; ICT with cues is the explicit independent variable; measures of understanding indicative; breadth of understanding of evidence; the ability to generate and support an argument; classroom context not typical (girls-only school)
Weight of evidence D: Stimuli (overall weight of evidence)	Medium

Lavoie DR (1999) Effects of emphasizing hypothetico-predictive reasoning within the science learning cycle on high school students' process skills and conceptual understandings in biology. *Journal of Research in Science Teaching* **36:** 1127-1147.	
Country of study	USA
Details of researchers	Researcher at Black Hills State University, Dakota and teacher/researchers
Name of programme	HPD-LC = hypothetico predictive discussion learning cycle
Age of learners	15 to 16
Type of study	Evaluation: researcher-manipulated
Aims of study	To examine the effects (in terms of teacher and student attitudes and their conceptual understanding and logical thinking abilities) of including a prediction/discussion phase in the learning cycle (exploration, term introduction, concept application) prior to the exploration phase
Summary of study design, including details of sample	A comparative evaluation trial in which the experiences, achievements and attitudes of students being taught in one way are compared with those being taught by the same teacher but with a different instructional process. Five grade 10 teacher/researchers each taught one HPD-LC and one LC class for a three-month semester. Classes were selected to be as similar as possible. Actual sample: 10 teachers and approximately 250 students
Methods used to collect data	• Daily logs kept by teachers • Observations by non participant university researcher • Videotapes • Pre-post intervention tests • Post intervention questionnaires to students and teacher/researchers
Data-collection instruments, including details of checks on reliability and validity	• The three pre- and post-intervention tests were processes of biological investigation test (Germann, 1989), group assessment of logical thinking test (Roadrangka, Yeany and Padilla, 1983) and conceptual understanding in biology test (developed by researchers). • The Likert-scale questionnaires were to assess attitude towards science, the learning cycle, peers, teacher/students, and the treatments. Teacher researcher questionnaire also posed short answer questions concerning the positives and negatives of the learning cycle strategy and tips for its improvement. • Checks on reliability: Established for the three pre-post tests. There is no reporting of reliability measures for the observations made or the questionnaires used. • Checks on validity: One of the pre-post tests (concept understanding) was subject to content validity checks. Not reported for observations or questionnaires.
Methods used to analyse data, including details of checks on reliability and validity	• Data from classroom observations and teacher/researcher daily observation logs were synthesised and categorised into coded statements, reflecting the researchers impressions of HPD-LC and LC instructions (reference to Bogdan and Bilken, 1982 = a reference text on qualitative research methods). Only those categorised observations that occurred within and between each class of the HDP-LC classes or the LC classes are reported. • Unpaired t-tests were used to compare pre-test scores for intervention and control groups to determine equivalence. • Unpaired t-tests were used to compare post-test scores for intervention and control groups to determine equivalence. • Paired t-tests were used to compare pre and post-test scores for control and intervention groups to determine equivalence within groups. • Mean scores and percentage responses in each category for the final teacher/researcher and student questionnaires are calculated.

	• Scores on the student questionnaires of the HPD-LC (intervention) and LC (control) groups were compared with the ci-square statistic. • Checks on reliability: No information is reported. • Checks on validity: No information is reported.
Summary of results	• Prediction/discussion-based learning cycle (HPD-LC) instruction compared with traditional learning cycle instruction produced significant gains in the use of process and logical thinking skills, science concepts and scientific attitudes. • In general, teachers felt that learning cycle instruction was more effective than their normal teaching mode for revealing students' misconceptions, teaching process skills and teaching some concepts. • Teacher/researchers were generally more satisfied with prediction/discussion-based learning cycle (HPD-LC) instruction than traditional learning cycle (LC) instruction and displayed a more positive attitude toward their the HPD-LC students. • Student questionnaire data revealed strong trends favouring learning cycle instruction.
Conclusions	• Both learning cycle instruction sequences (control and intervention) verify previous results that LC instruction improves reasoning skills, conceptual achievement and scientific attitudes. • The HPD-LC instruction (the intervention) compared with the traditional LC instruction achieved significantly greater gain scores for science process skills, logical thinking and conceptual understanding, and authors give four suggestion why this may be. • Positive outcomes of HPD-LC may only be achieved if teachers are trained in the application of learning cycle instruction; are willing to meet regularly to discuss their work; and attempt to standardise their instruction.
Weight of evidence A (trustworthiness in relation to study questions)	Medium There is convincing evidence in the test scores pre- and post-intervention. Possible bias is unresolved. There is mention that teacher/researchers displayed a more positive attitude towards the students receiving the intervention. The implications of this are not discussed.
Weight of evidence B: Stimuli (appropriateness of research design and analysis)	Medium-high Large sample size (250 students), but only one school; control group used for non-users of written materials; pre-post data collected for logical thinking; validity/reliability for data collection very detailed; validity/reliability for data analysis, t-tests for validity, little on reliability.
Weight of evidence C: Stimuli (relevance of focus of study to review)	Medium Written materials asking for written prediction are very commonly used as stimulus (print materials); stimulus is pretty central but combined with consensus building as independent variable; measures for understanding of evidence are limited to 'logical thinking'; breadth of understanding of evidence limited to engagement with data; whole class setting with small group prediction agreement is very representative.
Weight of evidence D: Stimuli (overall weight of evidence)	Medium

Palincsar AS, Anderson C, David YM (1993) Pursuing scientific literacy in the middle grades through collaborative problem solving. *Elementary School Journal* **93:** 643-658.	
Country of study	USA
Details of researchers	Researchers at the University of Michigan and the State University of Michigan
Name of programme	Collaborative Problem-Solving Program
Age of learners	11 to 12
Type of study	Evaluation: naturally-occurring
Aims of study	To evaluate the effects of an intervention including guidance of the use of scientific explanations and constructive group interaction on the ability to apply knowledge of kinetic molecular theory to everyday problems
Summary of study design, including details of sample	The collaborative problem-solving programme involved using a sequence of activities on kinetic molecular theory with nine grade 6 classes in two schools over a period of two years. Students were placed in groups of four, heterogeneous with regard to gender and race. Discussion tasks were aimed at modelling the working of scientific communities. A variety of data was collected (see later sections). This study focuses on analysis of discourse. Actual sample: Nine classes with an average of 26 students implies a sample size of around 230 students.
Methods used to collect data	• Curriculum-based assessment: pencil-and-paper tests of conceptual understanding • One to one interview • Observation: video recordings of particular groups • Self-completion report or diary: student logs
Data-collection instruments, including details of checks on reliability and validity	No details given
Methods used to analyse data, including details of checks on reliability and validity	• The use of a t-test for pre- and post-intervention results assumed. • Grounded theory seems to have been used for the analysis of group and class discussions. With comparison between year 1 and year 2 observations. • Checks on reliability: No details given, other than, by implication; multiple data sets enhance reliability. • Checks on validity: Triangulation between student logs and recorded group discussions forms some type of validity. Authors do not mention having done this.
Summary of results	• Students initially approach problem-solving very differently from adult scientists, in ways in which teachers would characterise as careless, immature or unthinking. This changed over time. • Poster presentations revealed contradictions in results, which in turn led to discussion of accuracy of reporting. • Students initially found whole class discussion and debate about reaching a consensus confusing, but did ultimately arrive at an agreed scientific view. • Students enjoyed planning the investigation. • Students used explanations to scaffold their discussions, particularly to provide reasons for their proposals. Students also discussed explanations. • Students stayed focused on discussion tasks.

	• Students were able to use their previous everyday experience to inform planning of investigations. Students demonstrated some of the characteristics of engaging in the enterprise and language of science, particularly in the second year of the study. • Post-test measure of understanding showed that a significantly greater number of students in year 2 achieved the targeted conceptual goal. • No significant difference in pre-test for year 1 and pre-test for year 2 [t(82) = 1.05, p = 0.296], but significant difference on the post-test [t(82) = 2.625, p = 0.005]. On the post-test 36.6% in year 1, and 51.1% in year 2 provide explanation for dissolving including both macro and micro-elements. 24.4% in year 1 and only 6.4% in year 2 provide naive responses.
Conclusions	Specific conclusions of the study are not summarised, but are implicit in the reporting of the data. The conclusions focus on teacher needs to support the use of activities such as those described in the paper.
Weight of evidence A: (trustworthiness in relation to study questions)	Medium The findings do seem trustworthy to a degree in that they seem sensible. The lack of detail on issues of validity and reliability reduces the trustworthiness of this study as reported here.
Weight of evidence B: (appropriateness of research design and analysis)	Medium-low Large same size; no control group, unless subsequent cohorts with different practical tasks are considered as such; no pre-intervention data for understanding of evidence (but for conceptual understanding); reliability/validity of data collection: no detail; reliability/validity of data analysis: no detail.
Weight of evidence C: (relevance of focus of study to review)	Medium Stimulus highly representative (standard practical on dissolving sugar); stimulus (practical) not sole but combined independent variables (together with scaffolding strategy); measures only descriptive; breadth of measure is average with focus on constructing models from information provided; setting (whole class) is only unusual as learners did not usually work in groups.
Weight of evidence D: (overall weight of evidence)	Medium

1. Sherman GP, Klein JD (1995a) The effects of cued interaction and ability grouping during cooperative computer-based science education. *Educational Technology Research and Development* **43:** 5-24. 2. Sherman GP, Klein JD (1995b) The effects of cued interaction and ability grouping during cooperative computer-based science education. Arizona, USA: ERIC report number ED 383769.	
Country of study	Assumed that study was in USA junior high school, probably in Arizona
Details of researchers	Researchers at Emporia State University and Arizona State University
Name of programme	Designing and Controlling Experiments (computer-based instructional program)
Age of learners	13 to 14
Type of study	Evaluation: researcher-manipulated
Aims of study	To investigate the effects, in terms of conceptual understanding, attitude and group behaviour, of verbal interaction cues and ability groupings within a co-operative CLE.
Summary of study design, including details of sample	Study was experimental with dyads of learners, grouped according to ability (high/high: low/low: high/low), working through a cued (for verbal interaction) or non-cued version of a CBI program on designing controlled experiments. Student performance on practice questions (answered in the dyads) and on a post-test (answered by individuals) was scored as was attitude to CBI and to working with each other. Behaviour while working on the CBI program was recorded. Actual sample: 256 students were initially involved, useful data were provided by 231.
Methods used to collect data	• Observation • Self-completion questionnaire
Data-collection instruments, including details of checks on reliability and validity	• Practice test items • Post-test items • Likert attitude scale • Video of interaction • Time records of computer work • Checks on reliability: KR21 reliability for post-test items; Cronbach Alpha for attitude • Checks on validity: no details provided
Methods used to analyse data, including details of checks on reliability and validity	• Scoring of attainment tests; statistical analysis of quantitative data from test; allocation of behaviour recorded on video into nine predetermined categories • ANOVA • MANOVA • Tukey HSD pair-wise comparison • Checks on reliability and validity: not stated
Summary of results	• Students using the cued version of the program performed significantly better on the post-test than students using the non-cued version. • Direct observation of students showed that students in cued dyads exhibited significantly more summarising and helping behaviours than non-cued students. • Higher ability dyads exhibited significantly less off-task behaviour than the other dyads.

Conclusions	• The main conclusion is that providing CBI with cues to encourage collaborative working does result in less off-task activity and improved test results.
Weight of evidence A: (trustworthiness in relation to study questions)	High A considerable weight of data that has been subjected to rigorous statistical analysis supports the conclusions of the study.
Weight of evidence B: (appropriateness of research design and analysis)	Medium Sample was large (N=231) and sampling method carefully done, but only from one school; control group included for stimulus (cued versus uncued ICT); pre-testing done only on conceptual understanding (not understanding of evidence) and solely for group composition; reliability of data collection high (Kuder-Richardson and Cronbach's alpha methods) but nothing on validity; reliability of data analysis by ANOVA statistical tests, but nothing on validity.
Weight of evidence C: (relevance of focus of study to review)	Medium Stimulus very representative for ICT; central focus (independent variable) of the intervention on cued versus uncued ICT stimulus; three measures (attitudes, reasoning skills and process skills) but none really specifically on understanding of evidence; breadth of measuring understanding of evidence minimal; setting representative in computer lab.
Weight of evidence D: (overall weight of evidence)	Medium-high

Suthers D, Weiner A (1995) Groupware for developing critical discussion skills. In: Schnase JL, Cunnius EL (eds) *Proceedings of CSCL 1995: The First International Conference on Computer Support for Collaborative Learning*. New Jersey, USA: Lawrence Erlbaum Associates Inc. pages 341-348.	
Country of study	USA
Details of researchers	Researchers at the University of Pittsburgh funded by a NSF Applications of Advanced Technology programme.
Name of programme	Belvedere software environment
Age of learners	15 to 16
Type of study	Evaluation: naturally-occurring
Aims of study	To undertake a formative evaluation of a specific CLE to stimulate collaborative formulation of a scientific argument, and thus to promote learning of science concepts and reasoning
Summary of study design, including details of sample	It uses a cross-sectional design for a formative evaluation with three cycles of prototype refinements. The first two result in interface refinements. The last action research cycle results in the extension of collaboration between students on one computer, to collaboration of students on two adjacent computers. Interactions of several collaborating dyads/triads were collected and tested in eight sessions in grade 10 classrooms in which students worked at computers. No longitudinal analysis done: The study intends to identify issues, not (causal) relationships. Actual sample: Two cycles of prototype testing with eight participants, individuals and dyads respectively The third prototype cycle with unspecified number of participants The main evaluation with unspecified number of participants
Methods used to collect data	• Observation: tape-recorded group conversations • Computer logs
Data-collection instruments, including details of checks on reliability and validity	• Not stated/unclear • Checks on reliability: none • Checks on validity: It is inferred that the experience used to collect the data in the three cycles of prototype development in themselves improve the validity of the strategy for collecting the data of the evaluation.
Methods used to analyse the data, including details of checks on reliability and validity	Not reported
Summary of results	• Belvedere facilitates the generation of several alternative hypotheses, forming the basis of argumentation. • Typically, more hypotheses were generated orally than entered in the Belvedere diagram. • Students use peer-coaching strategies within the groups to complement each others' content and IT knowledge. • Conflicting hypotheses cause (in some groups) fruitful dialectic tension between challenge and resistance to change proposed views. Subsequent debates, with scaffolding and reflection, provide personal experience of scientific dialectics. • Social (group) processes may preclude constructive participation and engagement with conflict. • Scientific argumentation skills require apprenticeship or practices not found in peer group. Further need for 'automated advisor' as part of the Belvedere package.

Conclusions	• Belvedere works. • There is a need to scaffold scientific argumentation skills in the software. • There is a need for further development of Belvedere to strengthen role of scaffolding 'automated advisor'.
Weight of evidence A: (trustworthiness in relation to study questions)	Medium-low Small sample with unclear composition strategy, lacks reliability checks on data-collection, and reliability and validity checks on data analysis.
Weight of evidence B: (appropriateness of research design and analysis)	Low Sample size small, with unclear sampling method; no comparison/control for stimulus; no pre-post data, although some action research data informed the next cycle; reliability/validity of data collection mainly focused on describing interaction with software features; reliability/validity of data analysis: little detail provided.
Weight of evidence C: (relevance of focus of study to review)	Medium Nature of ICT stimulus not very representative: complicated and of a trial nature; ICT stimulus (Belvedere) is the sole explicit independent variable; measures for understanding of evidence very vague; broad range of evidence reported (constructing argument on the basis of information provided); setting (two or three students working on same document on different computers) not typical.
Weight of evidence D: (overall weight of evidence)	Medium-low

Tao PK (2001) Developing understanding through confronting varying views: the case of solving qualitative physics problems. International Journal of Science Education **23**: 1201-1218.	
Country of study	China: Hong Kong
Details of researchers	University-based researcher working on funded project. A research assistant is also mentioned. There are some indications in the text to suggest elements of practitioner research or research undertaken for a higher degree, although no details are given.
Name of programme	No details given
Age of learners	17 to 18
Type of study	Evaluation: naturally-occurring
Aims of study	To explore whether and how group discussion of feedback of multiple alternative solutions to qualitative physics problems helped to improve students' problem-solving skills and understanding of underlying physics concepts
Summary of study design, including details of sample	A case study focusing on the evaluation of three qualitative physics problems The sample consisted of a convenience sample of one class of 18 year 12 students, of whom 16 were included in the analysis. The study involved four stages: a pre-test, feedback, a post-test (of three parallel questions similar to the three in the pre-test) and semi-structured interview. In the first two stages, students worked in dyads and their peer-interactions were audio-recorded. The post-test and interview involved individual students.
Methods used to collect data	• Curriculum-based assessment (physics problems) • Group interview • One-to-one interview • Audio tapes of discussion work
Data-collection instruments, including details of checks on reliability and validity	• Three qualitative problem tasks on mechanics, circuit electricity and optics for the pre-intervention task • Three (similar) qualitative problem tasks on the same topics for the post-intervention task • Example of various alternative solutions to problems for feedback phase • Semi-structured interview schedule • Checks on reliability: A research assistant also marked the students' responses on the pre-test; the use of three tasks intended to measure the same effect increases the reliability. • Checks on validity: No details are given of validation of interview schedule. • Validity of equivalence of pre- and post-intervention tests was improved as follows: use of pre-intervention test from previous study means the tasks have been piloted; a panel of three experienced physics teachers judged the parallel post-test questions to be comparable with the pre-test questions; validation of equivalence of level of difficulty of pre- and post-test by administering both tests to other class of 35 students, divided randomly, matched according to national exam results - results from pre-test taken by group 1, post-test taken by group 2 analysed by Mann-Whitney test show a mean score of 17.75 and 18.26 and p = 0.87. Validity of feedback instrument with varying alternative solutions is certain since actual student scripts have been copied to form the basis of this.
Methods used to analyse data, including details of checks on reliability and validity	• Problem-solving skills: no details given • Understanding of physics concepts: analysis of discussion, interview transcripts and students' written reflections on feedback sheet. • Frequencies • Statistics (Wilcoxon signed rank test) for analysing both pre- and post-test • Analysis of discussion, interview transcripts and students' written reflections on feedback sheet

	Wilcoxon signed rank test shows 4.33 for positive ranks (post test > pre test), two-tailed significance level p = 0.037; so improvement at 0.05 level. - Reliability of data analysis: Responses to pre-test for four random scripts (25%) were coded independently by two researchers with high agreement. - Validity of the data analysis was improved by triangulation of tape-recorded interactions, student scripts and interviews, and the use of a coding scheme used in a previous study.
Summary of results	- Students' understanding is enhanced and their problem-solving skills improved through the intervention. - Students valued the discussion tasks. - Students were generally positive about the process; three of the 18 expressed negative views. - Students were prompted to reflect on their approach to learning physics (metacognition).
Conclusions	The author concludes that the intervention offers exciting possibilities for developing students' conceptual understanding of physics, particularly through presenting students with multiple solutions to problems.
Weight of evidence A: (trustworthiness in relation to study questions)	Medium Indicators for problem-solving skills are not clearly stated. Reported abilities (e.g. meta-cognition) are unrelated. Reliability and validity of data-collection methods and analysis methods are not specified. The validity and reliability of data-collection method and analysis method is high. The research design could have included a control group. The small sample size causes some reservations about the generalisability.
Weight of evidence B: (appropriateness of research design and analysis)	Low Small sample (18 students from one school); no control/comparison group; no benchmark data on understanding of evidence (some on conceptual understanding); validity/reliability of data collection: equivalence test for pre- and post-test with Spearman-Brown split-half method, no reliability checks. Post-test was delayed 3.5 months, and done as individuals, whereas pre-test was done in groups; validity/reliability of data analysis: pre-post Wilcoxon signed rank test for significance directed to conceptual understanding, not understanding of evidence.
Weight of evidence C: (relevance of focus of study to review)	Medium Independent variable is specifically the written multiple correct problem solutions, reasonably representative; multiple correct problem solutions are sole/explicit independent variable; measures problem-solving skills, rather than understanding of evidence; breadth of understanding of evidence is minimal (see above); research setting is partly representative (in high-ability classrooms).
Weight of evidence D: (overall weight of evidence)	Medium-low

Tao PK (2003) Eliciting and developing junior secondary students' understanding of the nature of science through a peer collaboration instruction in science stories.. International Journal of Science Education **25**: 147-171.	
Country of study	China: Hong Kong
Details of researchers	University based researcher
Name of programme	None
Age of learners	11 to 16
Type of study	Evaluation: naturally-occurring
Aims of study	To elicit students' understandings of the nature of science (NOS) and investigating how students reacted to the science stories in a peer collaboration setting
Summary of study design, including details of sample	Pre-post test, no control. 150 students in four classes taught by the same teacher experienced the intervention over five lessons on the NOS. The intervention was based on stories about science. Students worked in pairs on these and answered pre-post test questions in pairs. The pre-post tests consisted of the same four multiple-choice questions on the NOS. One of the four classes, termed the focus class (18 pairs), were video and audio-taped and their interactions analysed. Two months after the intervention, individuals from nine of the pairs were interviewed. These were chosen randomly from groups showing no change in their understanding, improved understanding of NOS and deterioration in their understanding.
Methods used to collect data	• Pre-post tests for whole sample of 150 students • Video and audio tapes for 36 students in the focus class • Field notes for observations of the focus class • Interviews with 18 students from the focus class
Data-collection instruments, including details of checks on reliability and validity	• Four multiple-choice questions, adapted from a test published by Solomon *et al.* (1996), were used for the pre-post test on the NOS. • The questions were mapped against three aspects of the NOS. • Questions were confirmed as appropriate by five experienced science teachers. • Observations, using video and audio recording and field notes, provided reliability for student interactions. • Interviews, using randomly selected students of three different achievement levels, provided further checks on reliability and validity of the data collected.
Methods used to analyse data, including details of checks on reliability and validity	• Data from four classes (150) students were presented as % for each of the four questions but no statistical test was applied. • The process of testing assertions against data together with the description of the 'natural history of enquiry' was seen to support claims for the credibility of the results. • Reliability analysis comes from the procedure of accepting or rejecting an assertion on the basis of accumulated evidence in the data sources. • For pairs in the focus class, responses to post-test items were cross-checked against transcripts of their interactions.
Summary of results	• Many students have entrenched inadequate views of NOS. • Students can give articulate and sophisticated arguments, irrespective of whether these views are adequate or not. They draw on prior knowledge and/or science stories for such arguments. • The science stories in the NOS instruction influence students in substantial ways but not always to improve understanding. • When studying the science stories, many students selectively attend to certain aspects that appear to confirm their inadequate views of NOS. They are unaware of the overall theme of the stories as intended by the instruction. • The peer collaboration provided students with experiences of conflict and co-construction that helped them develop shared

	understanding of NOS. However many students interpreted the science stories in idiosyncratic ways other than as intended by the instruction and subsequently changed from one set of inadequate views of NOS to another rather than to adequate views.
Conclusions	• Science stories influence students' views on the NOS. • Science stories provide useful contexts for students to offer arguments in support of their views of NOS. • Science stories used in a peer collaboration setting can be a form of argument-based pedagogy (in contrast to teacher dominated discourse). • The limited success of the NOS instruction may be attributed to the deeprootedness of students' inadequate views of NOS and the short duration of the instruction or may be due to the ways students make sense of these stories that differ from those intended by the instruction. However, it is argued that the problem arises from students' construction of meanings and sense-making of the stories, a finding that is consistent with constructivists' view of learning in which students' prior knowledge plays an important part. • The peer-collaboration setting did not help the situation much. The conflict and co-construction arising from the collaboration could lead to adequate as well as inadequate views of NOS. • Without guidance from the teacher students tend to have idiosyncratic interpretations of stories that match their inadequate views of NOS.
Weight of evidence A: (trustworthiness in relation to study questions)	High This is a carefully planned and executed study with strong checks for reliability and validity for data collection and analysis. Care is taken to compare the results of post-test findings of pairs of students with their own pre-test results rather than with the class average. The results and conclusions address the research questions and discuss the several findings in relation to possible pedagogical explanations.
Weight of evidence B: (appropriateness of research design and analysis)	Medium-high Good sample size for all three stages of the research and sampling frame; no comparison for stimuli; good benchmark information from pre-test; solid checks for reliability and validity of data collection; solid checks on reliability and validity of data analysis.
Weight of evidence C: (relevance of focus of study to review)	Medium-high Type of stimulus (stories) is used but not traditional or widespread; major focus of intervention is the stimulus; measures are highly appropriate for testing understanding; good breadth of measures for understanding of science; situation is only representative of classroom learning for higher ability boys.
Weight of evidence D: (overall weight of evidence)	Medium-high

Williams A (1995) Long-distance collaboration: a case-study of science teaching and learning. In: Spiegal SA (ed.) *Perspectives from Teachers' Classrooms. Action Research. Science FEAT (Science for Early Adolescence Teachers).* Tallahassee, FL, USA: Southeastern Regional Vision for Education.	
Country of study	USA
Details of researchers	One practitioner researcher, part of a project which appears to have been co-ordinated by Stanford University
Name of programme	Human Biology Middle Grades Life Science Curriculum Development Project (HumBio)
Age of learners	11 to 12
Type of study	Evaluation: naturally-occurring
Aims of study	Not very specific, but to assess the benefits to students of a project (on abiotic and biotic materials used in modelling an environment), completed in collaboration with a distant school
Summary of study design, including details of sample	A case study of the implementation of one of the three curriculum intervention packages developed for the HumBio project and used with three classes (90-100 students), all taught by the researcher. The activities and views of students in three classes were recorded as they provided and exchanged materials with a distant school. Data were gathered from the one school (Florida). Students working in small-groups tried to map the grounds of the distant school from the information provided and groups compared maps. They then did their own survey and finally watched a video from the distant school to compare it with their own mapping of that school. Written feedback was collected from the students. Hard copies of email correspondence and students' work were kept. Videotapes were made of selected group activities. Field notes were kept by the researcher.
Methods used to collect data	• Observation: video recordings of student group presentations • Self-completion questionnaire • Teacher notes and journal • Other documentation: (a) students' work including drawings and models; (b) email correspondence with another school; (c) written reflections solicited after each of the three stages of the project
Data-collection instruments, including details of checks on reliability and validity	• Questionnaire; no details of reliability or validity checks
Methods used to analyse data, including details of checks on reliability and validity	• No details are given and there was no formal analysis. Quotes from students are included to support conclusions. • No details of reliability or validity checks other than, by implication, the notion that the use of multiple data sources enhances validity.
Summary of results	• Students have fun while learning. • Project makes learning more relevant and meaningful to students by providing a practical 'real world' purpose for the learning experience. • It provides practice in the use of science process skills. • It extends co-operation and collaboration among students by expanding the field of interaction beyond the classroom, school and state.

	• It fosters the development of the scientific attitudes of imagination, openness to new ideas and scepticism.
Conclusions	• Difficulties with electronic communication (asynchronous discussions) suggest that participants need to agree a schedule for routine checking and replying. • Participation in intervention helps students' metacognitive processes. • Despite problems associated with timing and co-ordination, the collaboration provided a relevant and meaningful learning experience which students found enjoyable. Students received practice in the use of science process skills, as well as scientific attitudes of imagination, openness to new ideas and scepticism. • The project could serve as a good model for scientific enterprise by providing students with the experience of doing what scientists do.
Weight of evidence A: (trustworthiness in relation to study questions)	Low Little data are reported on the small-group discussion work. The study is a small-scale evaluation, undertaken by an enthusiastic teacher and reported in what looks like a practitioner journal. Data have not been formally analysed.
Weight of evidence B: (appropriateness of research design and analysis)	Low Small sample; no control group for stimulus (practical work); only post-intervention data collected; reliability/validity of data collection: no checks; reliability/validity of data analysis: no checks.
Weight of evidence C: (relevance of focus of study to review)	Medium-low Practical work for networking software with other schools not representative for practical work stimulus; stimulus central independent variable; measures for understanding of evidence obscure; breadth of understanding of evidence lacks information; practical work is standard but data exchange with other schools not representative of practical work.
Weight of evidence D: (overall weight of evidence)	Low

Zohar A, Nemet F (2002) Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching* **39:** 35-62.	
Country of study	Israel
Details of researchers	Two university-based researchers; some of the data appear to have been collected by teachers
Name of programme	Thinking in Science Classrooms: Genetic Revolution unit
Age of learners	13 to 14 (age not specified, but described as 'grade 9')
Type of study	Evaluation: researcher-manipulated
Aims of study	To examine the effects of a unit that teaches argumentation skills in the context of dilemmas in human genetics, focusing on development of biological understanding and argumentation skills
Summary of study design, including details of sample	186 participants in two schools were assigned to a control group (99 students, five class sets) and an experimental group (87 students, four class sets). The assignment of classes to experimental and control groups was random. The experimental group received the Genetic Revolution unit, which took twelve lesson of teaching time. It is not immediately clear how many teachers were involved. The implication is eight, of which three taught both a control and an experimental group. Each group received a pre- and post-test of argumentation skills and biological knowledge. A multiple-choice test, audiotaped discussions and written worksheets were used to gather data. Actual sample: Not all students were included in the analysis, due to absence when some of the data were collected. No details of the final samples size are given.
Methods used to collect data	• Curriculum-based assessment: 20 multiple-choice items • Student worksheets • Audiotapes of four small-group discussions
Data-collection instruments, including details of checks on reliability and validity	• 20 multiple-choice items to assess biological knowledge • Worksheets to assess argumentation skills • Audiotapes of four small-group discussions • Checks on reliability: No details about reliability are given. • Checks on validity: Some of the multiple-choice items were from previous years' examinations and some developed for the study, with the content validity of the latter items being checked by an expert.
Methods used to analyse data, including details of checks on reliability and validity	• Qualitative categories based on previous research were used in analysis of audiotaped discussions. • Researcher-developed method to score pre- and post-tests of argumentation skills • Calculation of inter-rater reliability scores for argumentation analysis • t-test of significance of use of biological knowledge in post-test • t-test of significance of mean scores on argumentation tests • Test of 'frequency of conclusions' • Checks on reliability: Argumentation skills analysis was done by both researchers, and inter-rater reliability scores calculated. • Checks on validity: No details are given.

Summary of results	• Following instruction, the number of students using correct, specific biological knowledge in constructing arguments increased from 16.2% to 53.2%. • Students in the experimental group scored significantly higher than students in the control group in a test of genetics knowledge. • Analysis of the written tasks showed an increase in the number of justifications and in the complexity of argument. • Students were able to transfer reasoning abilities tools in the context of bioethical dilemmas to the context of dilemmas taken from everyday life. • There were dramatic changes in the quality of student arguments. • Changes were detected in the frequency of explicit conclusions, the mean number of justifications for a conclusion and in the number of ideas students expressed while talking. • Integrating explicit teaching of argumentation into the teaching of dilemmas in human genetics enhances performance in both biological knowledge and argumentation.
Conclusions	• Students showed improved understanding of biological concepts. • Teaching through social issues provides 'anchored instruction' for students by generating interest and connecting to out-of-school life experiences. • Student learning was aided by having students work in small groups for substantial amount of time in most lessons. • Argumentation skills were enhanced by explicit instruction about the formal structure of an argument, and the generation of multiple opportunities for students to take part in discussions that require intensive use of arguments. • Reasoning about dilemmas should be integrated into other science topics. • The authors advise caution against making unsupported generalisations from their findings as they suggest that many may relate to specific properties of the context of the intervention. They also note that many of the teachers and students were very enthusiastic about the programme, again suggesting caution against over-generalising from the findings.
Weight of evidence A (trustworthiness in relation to study questions)	Medium Possible researcher and teacher bias mean that the findings have to be treated with some caution. No details are given of how schools and teachers were recruited into the study.
Weight of evidence B (appropriateness of research design and analysis)	Medium-high RQs 4 & 5 relevant to this review. Sizeable sample size, but little detail provided on the sampling method; use of control group; pre-post data collected on argumentation skills; reliability/validity on data collection: no detail provided apart from content validity check of test items by peer expert; reliability/validity on data analysis: ANOVA statistics and inter-rater coding ratios provided.
Weight of evidence C (relevance of focus of study to review)	Medium-high Nature of stimulus (prepared curriculum materials) very representative; stimulus together with training in group discussion techniques constitute the independent variable; measures (tests and discussion transcripts) appropriate for testing understanding of evidence; focus of understanding of evidence on argumentation skills; classroom setting representative.
Weight of evidence D (overall weight of evidence)	Medium-high