REVIEW

May 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

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 Judith Bennett
Department of Educational Studies
University of York
York YO10 5DD

Tel: 01904 433471
Email: jmb20@york.ac.uk

AUTHORS AND REVIEW TEAM

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

Database management and administrative support

Alison Robinson, Department of Educational Studies, University of York, UK

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 ABREVIATIONS

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
HSD	Honest Significant Differences

GALT	Group Assessment of Logical Thinking
GCSE	General Certificate for Secondary Education
ILL	Inter-Library Loan
MANOVA	Multi-Analysis of Variance
PGCE	Post Graduate 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
UYSEG	University of York Science Education Group
VOSTS	Views On Science Technology and Society

This report should be cited as: 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.

TABLE OF CONTENTS

SUMMARY

Background

This review focuses on 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 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 (Department for Education and Employment (DfEE), 1998) in England and Wales)

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 review research question 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?

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 mapping of the area revealed a wide range of relevant studies. A more limited focus was therefore adopted for the in-depth review, with the review question being limited to evaluative studies of students' understanding of evidence in science.

The in-depth review research question is:
What is the evidence from evaluative studies of the effects of small-group discussions on 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).

Methods

The review methods 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 compiling the weighted findings of the collective studies.

Results

The number of studies identified through the searching and screening processes established that small-group discussions were being used in a variety of ways in science lessons. However, a characteristic of many of the studies was that small-group discussions in themselves were rarely seen as discrete independent variables for investigation. Rather, the notion of small-group discussions tended to be wrapped up within other activities, often characterised as 'collaborative learning', a term which was used in a variety of ways and often very loosely, such that it appeared to include most activities which did not involve teacher exposition. This

resulted in a considerable amount of effort being required to refine searching, screening and keywording strategies to ensure studies fell within the review focus. Eighty-nine studies were identified for inclusion in the systematic map. The map revealed a number of characteristics of research on small-group discussions, as summarised below:

- The majority of the studies report work that has taken place in the US, the UK and Canada.
- Small-group discussions are used with all ages of student in the secondary age range.
- The majority of work focuses on small-group discussions in relation to students' understanding.
- A diversity of measures is 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 involve groups of three to four students emerging from friendship ties, and have a duration of at least 30 minutes.
- Typical small-group discussions have individual sense-making as their main aim (as opposed to, for example, leading to a group presentation) and use prepared printed materials as the stimulus for discussion.
- The most common research strategy was that of case study.
- Twenty-eight studies had experimental designs, of which 12 were randomised controlled trials (RCTs).
- The most popular techniques for gathering data are observation, videotapes and audiotapes of discussions, interviews, questionnaires and test results.

Fourteen studies were included in the in-depth review, which focused on the effects of small-group discussions on students' understanding of evidence in science.

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

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

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

- The use of small-group discussions based on a combination of internal conflict (i.e. where a diversity of views and/or understanding are represented within a group) and external conflict (where an external stimulus presents a group with conflicting views) resulted in a significant improvement of students' understanding of evidence. (From one medium-high rated study.)

- Improvement of students' understanding of evidence was not significantly different for members of all-female, all-male or mixed gender pairs. The benefit was greatest for female students when they were given several opportunities to engage with aspects of tasks related to understanding of evidence. (From one medium-high rated study.)

- Improvement of students' understanding of evidence correlated with the initial *dissimilarity* of the group members in terms of their domain-specific understandings; that is, student groups were constructed such that they contained students with as wide a range of domain-specific understandings as was possible. (From one medium-high and one medium rated study.)

- The use of small-group discussions did not affect students' ability to differentiate between observational or experimental data from opinions in a science-based text. (From one high rated study.)

- The use of small-group discussions supported by a specific programme fostering collaborative reasoning (including evaluating and strengthening of knowledge claims) improved students' metacognitive knowledge of collaborative reasoning (including their knowledge of reasoning about evidence) significantly more than for students not following the special programme. However, such gain within the treatment group depended on learners' perspective on learning: students with a *learner-as-explorer* perspective gained significantly more than peers with a *learner-as-student* perspective. (From one medium-high rated study.)

- The improved metacognitive knowledge of collaborative reasoning described above did not translate into better use of strategies while reasoning, including when dealing with scientific evidence. (From one medium-high rated study.)

The review suggests there is *some* evidence that:

- The use of either internal conflict small-group discussions (from one medium rated study) *or* external conflict small-group discussions (from one high and one medium rated study) produced improvement in students' understanding.

- The use of small-group discussions (together with specific instruction in argumentation skills) improved students' ability to construct more complex arguments. (From one medium rated study)

- The effectiveness of small-group discussions in producing an improvement in students' understanding of evidence depended on three types of understanding: understanding of the science domain, the process by which model-revision takes place, and metacognition. (From one medium rated study)

- The use of small-group discussions resulted in a significantly higher achievement in understanding of evidence for students using a cued version (that is, one which gives students specific instructions on what to include in points they make in discussions) of a computer-based instruction (CBI) program compared to a non-cued version. (From two medium rated studies.)

Although outside the specific focus of the in-depth review question, one additional finding worth noting is that there was reasonable evidence to suggest that the gender composition of small discussion groups determined the interaction style for developing students' explanatory understanding: all-male groups confronted differences in their individual predictions and explanations, while all-female groups searched for common features of their predictions and explanations across tasks, and mixed groups secured progress through turn-taking. (From one medium-high and one medium rated study.)

Conclusions

Strengths of the review

- 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.

- 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

- There was a scarcity of studies that focused on small-group discussions 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; that is, had an overall weight of evidence of medium or higher.

- Although the studies in the in-depth review shared a number of similar characteristics at the broad level, there were considerable differences at the detailed level. For example, 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 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 Team 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 small-group discussions adversely affect students' understanding of the nature of evidence. Therefore there is nothing to suggest that current policy (which is strongly advocating the use of small-group discussion work) should be changed. 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.

Implications for practice

The Review Team 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 teachers should 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 in providing a broader picture of research findings on small-group discussion work. Such areas would include the following:

- the nature of the stimulus provided for the group and its effect on the development of understanding;
- 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. argument, decision-making);
- the use of ICT in small-group discussions.

The Review Group will explore some of these areas in its next review.

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

The aim of the review is to focus on the ways in which small-group discussions are used in science lessons and their effects on students' understanding in science and attitudes to science. 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.

Specifically, the review research question 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?

As a result of a large number of studies being identified which addressed this question, a narrower focus has been taken for the in-depth review, where the research review question is:
What is the evidence from evaluative studies of the effects of 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 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 justifying 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 14-16-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* has recently begun its pilot phase in schools (September 2003), the outcomes of which 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, whilst 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 will 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 17-18 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. Whilst 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. 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 point very strongly to the desirability of a systematic review of the use of small-group discussions in science teaching. No systematic reviews for relevant topics currently exist.

A note on collaborative learning

There is a large quantity of mainly US-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 all the home groups is assigned a different portion of the material. The home groups then break apart and reform into 'expert' groups in which group members are all focusing on and discussing the same piece of the material to make sure they understand it. Once this has happened, students' groups then break once again and reform 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 reform, 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 rest of the 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.

Several members of the Review Group are also parents and a number are school governors (Judith Bennett, Martin Braund and Bob Campbell), and therefore represent additional constituencies with an interest in the work of the Review Group.

2.1.2 Methods used

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

- the decision over the review question;
- the development of inclusion and exclusion criteria;
- the development of review-specific keywords;
- the identification of the focus for the in-depth review;
- the content of the report.

The Review Group has met three times during the twelve-month period of the review to monitor and discuss progress, and to advise and guide the core team.

For example, during one of the meetings, the EPPI generic keywording sheet (EPPI-Centre, 2002a) was used by pairs of participants to code a paper circulated prior to the meeting. After plenary discussion of the experience, small groups of participants were asked to help construct the review-specific keywording categories by generating characteristics of studies on small-group discussions they wished to be documented, specifying for which key-users of the review these would be important. As a result, several unexpected aspects were included. Teachers' and curriculum designers' interest in synchronous and asynchronous group discussions were documented, as were the teacher educators' focus on group organisational aspects, such as snowballing, jigsawing and envoying.

In a subsequent meeting, the map of the various studies on small-group discussions was presented for comment. As a result, the report's presentation of the map information was re-structured. Subsequently, small-groups of participants were provided with cards with 14 potential in-depth review areas, each including manageable numbers of studies in the map. Groups were asked to prioritise these in-depth review areas, resulting in consensus on four distinct, but related, in-depth review areas. The most popular area was adopted for this report's in-depth review and

two of the remaining three have been agreed as areas for further in-depth reviews.

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 have 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 has been secured to involve Post Graduate Certificate in Education (PGCE) students in producing user-friendly summaries of the review findings for teachers, teacher educators and students. This will be part of their regular training programme. At the same time, the product will be 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 benefits 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 is to ensure that the review has 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 combination of papers reporting the same study.

2.2.1 Defining relevant studies: inclusion and exclusion criteria

The EPPI-Centre systematic review methods have been applied as described in the protocol. Thus the methods specified in this section have been followed for searching, screening and including (or excluding) studies in the map, and in applying

the EPPI keywording sheet and keywording strategy (EPPI-Centre, 2002a, 2002b), supplemented by review-specific keywords, to these studies. The review includes a descriptive mapping (identification and broad characterisation of the studies overall) prior to in-depth reviewing.

Studies are *included* in the systematic map if they satisfy the following criteria:

- They are about the use of small-group discussions in science lessons.
- They involve groups of two to six students.
- They involve a substantive, structured discussion task of more than one or two minutes duration.
- They illustrate how small-group discussions are being used.
- If focused on learning outcomes, they address aspects of students' understanding in science or they address aspects of students' attitudes to science.
- They are empirical studies of the following types: descriptive, exploration of relationships, evaluation (naturally-occurring and researcher-manipulated), reviews (systematic and non-systematic).
- They are about students in the 11-18 age range.
- They have been undertaken in the period 1980-2002.
- They are published in English.

Justification for focus on small-group discussions in science lessons

It is recognised that small-group discussions are used as a teaching strategy in several other school subjects, particularly in the humanities subjects and in personal and social education (PSE) programmes in schools. Although this review may benefit from reports of exemplary and effective practice in these other areas, the intended outcomes in science education are sufficiently distinct to justify a separate review.

Justification for size of group to be included

The review includes studies of groups of two to six learners for whom a substantive structured discussion task is set. Although usual group sizes in science classes typically vary from two to four learners, some tasks (such as, for instance, role plays) will require larger size groups. Groups in excess of six pupils would not be seen as *small group*s.

Justification for duration of discussion task to be included

There are instances in science lessons where students are given very short discussion tasks to stimulate their thinking, such as, for example, to talk to the person or people sitting next to them and agree on an answer to a question. Such discussion groups might be termed 'buzz groups'. The review seeks to gather information on more substantive discussion tasks associated with, for example, comprehension of science-based text; critiquing newspaper articles; planning investigations; interpreting science data; exploration of ideas and understanding; role plays; poster preparation; raising questions; and peer review.

Justification for types of study to be included

Specific types of studies are included in the review for the reasons listed below:

- descriptive – to yield information on the ways in which small-group discussions are used in science lessons;
- exploration of relationships – to yield information on relationships between the use of small-group discussions and, for example, students' performance in tests and examinations, or attitudes to science;

- evaluation – to yield information on the effects of small-group discussions in terms of what appears to be working, how it appears to be working and why;
- review – to gather information from reports which have attempted to bring together findings from a range of previous studies on the use of such approaches, whether systematic or non-systematic.

Justification for age range to be included

The principal groups with whom small-group discussion tasks are used in school science are students in the secondary and pre-university age range. Thus the age range covered by the review is 11-18.

It is recognised that there may be some merit for a subsequent review in looking at studies of work with primary age pupils, where discussion work may well be used more extensively than at secondary level.

Justification for period to be covered

Small-group discussion work in science has its origins in the constructivist research which began in the 1980s and the drive to broaden the range of teaching activities used in science lessons associated with the curriculum development work of the same period. Thus the period covered in the review is 1980-2002.

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.

Detailed formulation of inclusion and exclusion criteria is contained in Appendix 2.1.

2.2.2 Identification of potential studies: search strategy

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 1980-2002.
8 They are published in English.

Search strings are, therefore, be of the type:
1a and 1b and 2 and 3a or 3b and (4a or 4b) and (5a or 5b or 5c) and 6a and 6b and 7 and 8.

Since, searching on criteria 2, 3a and 3b was not practical (titles, abstracts or headings do not provide information on group size or the nature and duration of the task), these criteria were used only from the second stage screening onwards. Criterion 6a (about students) is implicit in the limits set for 6b (age range of 11-18).

The search string was restructured as:
1a and 1b and (4a or 4b), then limited for (5a or 5b or 5c) and 6b and 7 and 8.

Appendix 2.2 gives details of the search strategy terms used for electronic databases.

Reports were identified from the following sources:

- electronic databases: Educational Resources Information Centre (ERIC), PsycINFO, Social Sciences Citation Index (SSCI) and the British Education Index (BEI);
- search of journal publishers' web pages or handsearching of key journals for the period 1980-2002 (Appendix 2.3 gives details of these journals);
- citation searches of key authors/papers;
- reference lists of key authors/papers;
- references on key websites;
- personal contacts;
- direct requests to key informants.

2.2.3 Screening studies: applying inclusion and exclusion criteria

The Review Team 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 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 would be included on 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 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:

- 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

One step introduced into the review process was the identification at an early stage of a list of 'seminal papers' (SPs). These were papers that the core team felt, from their knowledge of the area, should emerge from the search stage. The team felt that this provided an essential check on the appropriateness of search terms and the thoroughness of the search as a whole.

The application of inclusion and exclusion criteria was initially conducted by all four core team members for a 2.5% random sample (45 papers). This was done independently in the first instance and the team members then met to compare the codes allocated, discuss the discrepancies and reach a consensus on how criteria were to be interpreted and applied. This enabled the clarification and removal of any ambiguities in their perceptions of coverage of the criteria. Only minor refinements of the exclusion descriptors were found to be necessary. Three team members were responsible for screening the remaining studies on the basis of abstract and title. These team members and a member of the EPPI-Centre worked on a second 2.5% random sample (45 papers), working independently. These data were used to calculate inter-screener agreement, using frequency counts and the Cohen's Kappa inter-screener reliability coefficients.

The strategy for screening on abstract and title was to 'include when in doubt' and request full papers through inter-library loans (ILLs). This involved a large number of papers purporting to report on co-operative and collaborative learning, and on small-group work. During the second stage screening of this cluster of full papers, the application of inclusion/exclusion criteria had to be refined considerably (see Table 2.1) in order to establish consistency in what counted as 'small-group discussion'.

Table 2.1: Application of inclusion/exclusion criteria for papers reporting on co-operative and collaborative learning (small-group work)

Description of intervention	General description of *structure* of group work: • No task (exclude 2) • Group work task (exclude 2) • Discussion task (include) Detailed description of *structure* of group work (no group discussion): • No task (exclude 2) • Group work task (exclude 2) Detailed description of *structure* of group work (including group discussion): • No task (exclude 4) • Group work/discussion task (include)
Description of results	• Discussion reported (include)

Table 2.1 shows that, if a paper provides a general description of the structure of group work or even a detailed description mentioning the use of group discussions, it was only included in the review if, in addition, at least one discussion task was described. Alternatively, papers were excluded on criteria 2 or 4 as indicated above.

Once the 114 papers (89 studies) to be included in the review had been identified, a 9% purposeful sample of ten papers with a variety of designs and focus was keyworded by all four core team members to check the appropriateness of the review-specific keywords and reach a consensus on how keywords were to be applied. Again, the team first worked independently and then met to compare keywording, discuss the discrepancies and potential changes to the review-specific keywords, and reach a consensus on how keywords were to be interpreted and applied. Following two rounds of amplification and modification to the review-specific keywords, thus increasing the validity, all papers were keyworded by at least two Review Group members or EPPI colleagues. In addition, a random 10% sample (11 papers) was keyworded independently by three core team members, together with a member of the EPPI-Centre.

The reliability has been increased by checking for logical conflict of keyword entry on the database. The validity of the keywording process was increased by checking the consistency per keyword category across clusters of similar papers.

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

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 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 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 does not focus on the effect of small-group discussions).
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 data-extraction procedures identify three categories - 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 categories as follows: high, medium-high, medium, medium-low and low.

The categories are as follows:

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 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 Team 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: the nature of the sample; the focus of the intervention; the appropriateness of the measures; the breadth of understanding of evidence measured; and the representativeness of the study situation. 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 Team 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 Team. Only one characteristic considered important was not included in this tool: the 'details of the researchers'. These reports were edited by one team member for consistency of terminology, depth and detail, continuously referring to each relevant study. The reports were used by two team members to identify commonalities across the studies for the same characteristics as presented in the map. In addition, communalities 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 team 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 primarily on the findings from studies weighted as *high*, *medium-high* and, to a lesser extent, as *medium*, as summarised above. Findings from studies weighted as *medium-low* are only considered if these corroborate findings of studies with a higher weight of evidence.

2.3.5 In-depth review: quality assurance process

A number of steps were followed in the review process for the purposes of quality assurance. Three core team members undertook a data-extraction on one of the in-depth review studies, working first individually and then meeting to moderate their summaries. This process increased the reliability of the subsequent data-extractions. Data-extraction of the remaining 13 studies was then conducted by changeable pairs of four core team members, working first independently and then comparing their decisions and coming to a consensus. In addition, for purposes of quality assurance, one member of the EPPI-Centre double data-extracted and quality assessed 35% (five) of the studies included in the in-depth review.

All members of the Review Team discussed and agreed the final decisions about weightings. Part of these discussions involved the consistency of the application of criteria for judgements of quality across the in-depth review as a whole. Another key aspect of this moderation involved developing the algorithm described in section 2.3.3 to assign weightings to specific aspects of the studies. The purpose of the algorithm was to give appropriate weightings to studies in each of the categories B and C, and to discriminate adequately between studies.

3. IDENTIFYING AND DESCRIBING STUDIES: RESULTS

3.1 Studies included from searching and screening

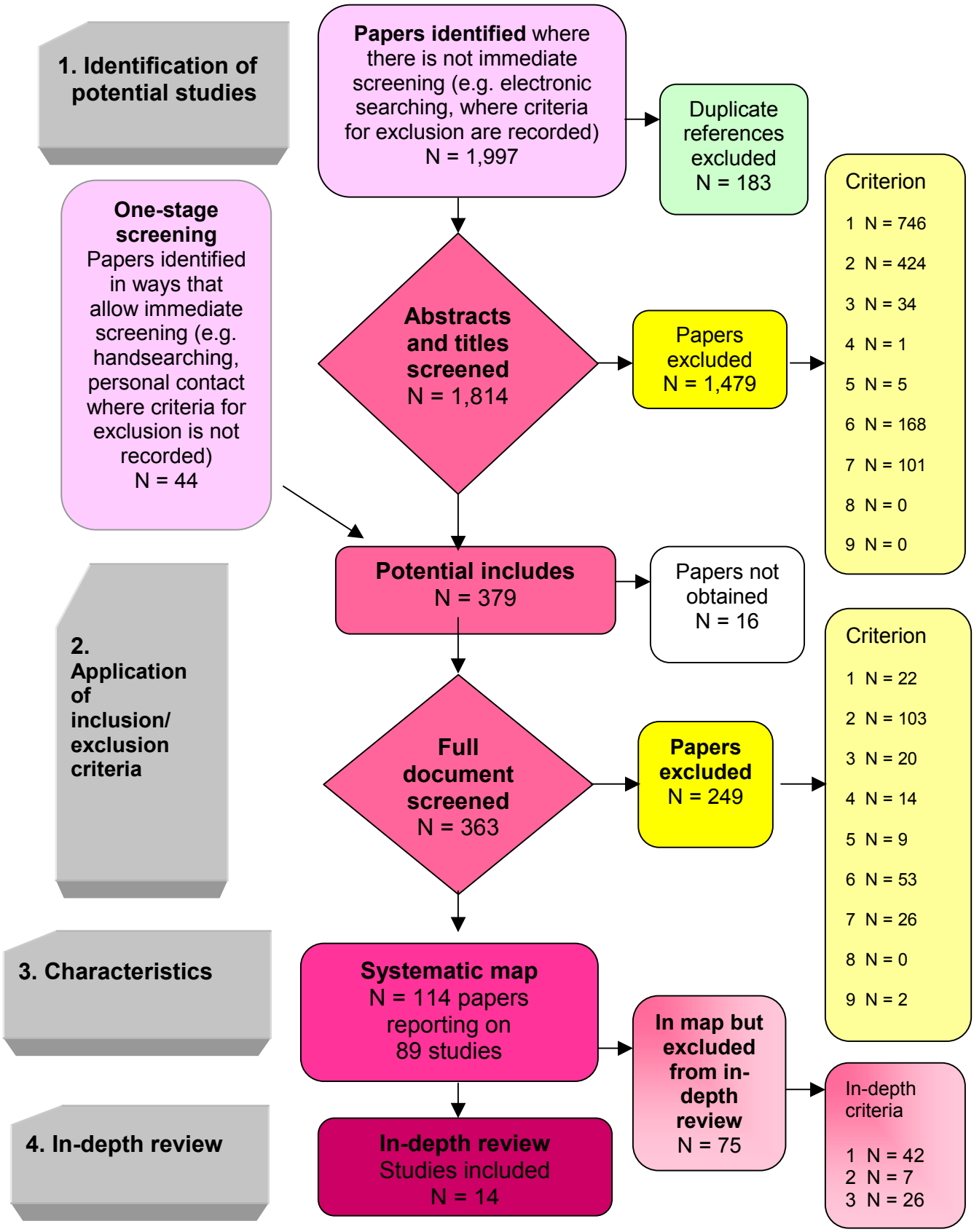
Figure 3.1 provides a summary of the number of papers and studies involved at various stages of the filtering process. The process of searching yielded 1,997 papers. An additional 44 papers were identified through handsearching or personal contacts; thus the review handled a total of 2,041 records. After de-duplication and the first round of screening 379 papers remained as potential includes. Hardcopies of only 16 papers (4%) were unobtainable. After second screening, 114 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 114 papers were found to report on 89 studies, 14 of which were included in the in-depth review.

Table 3.1: Origin of studies included in the systematic map

Source	Papers found	Papers included in map	Yield %	Studies in map from single source	Yield % of studies from single source (N = 89)
ERIC	836	60	7	16	18
BEI	56	19	34	2	2
PsycINFO	537	28	5	2	2
SSCI	568	58	10	17	19
Handsearch	39	6	15	5	6
Contact	5	3	60	3	3
Total	***2,041***	***174***		***45***	***50***

The first three columns of Table 3.1 show that the search of the ERIC, SSCI and PsycINFO databases generated over 500 papers each with a similar proportion of between 5% and 10% of papers included in the map for each of these databases. In other words, searches of all three databases have an equal, but very low, efficiency, with SSCI having a marginally higher yield percentage. The search of BEI, the handsearch and contacts each generated considerably less papers, but all with a higher efficiency. It is not surprising that the percentage yield for papers obtained through contacts (including through members of the Review Group) was very high (60%). On the other hand, only six papers (15%) emerging from the handsearch were eventually included in the map. This low percentage yield may be explained by the fact that several of these papers were key conference papers identified from a book of abstracts and papers cited in summary articles in *Studies in Science Education*, many of which could not be obtained.

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



The last two columns of Table 3.1 reflect the exclusiveness of the different databases in the identification of studies included in the map. They show that almost half the studies (49%) were represented in more than one of the databases, usually ERIC and SSCI. Slightly less than one in five studies in the map originated exclusively from each of ERIC and SSCI. A very low percentage (four studies) emerged solely from BEI or PsycINFO. These four studies included three studies reported in the *International Journal of Science Education* (a journal notoriously under-catalogued in ERIC) and one book chapter. Further inspection shows that the journal papers are all included in SSCI, although the current search strategy did not identify these three. This seems to indicate that, with a better search strategy for SSCI, there is no need to search the BEI and PsycINFO databases, as this will have only marginal returns, if any.

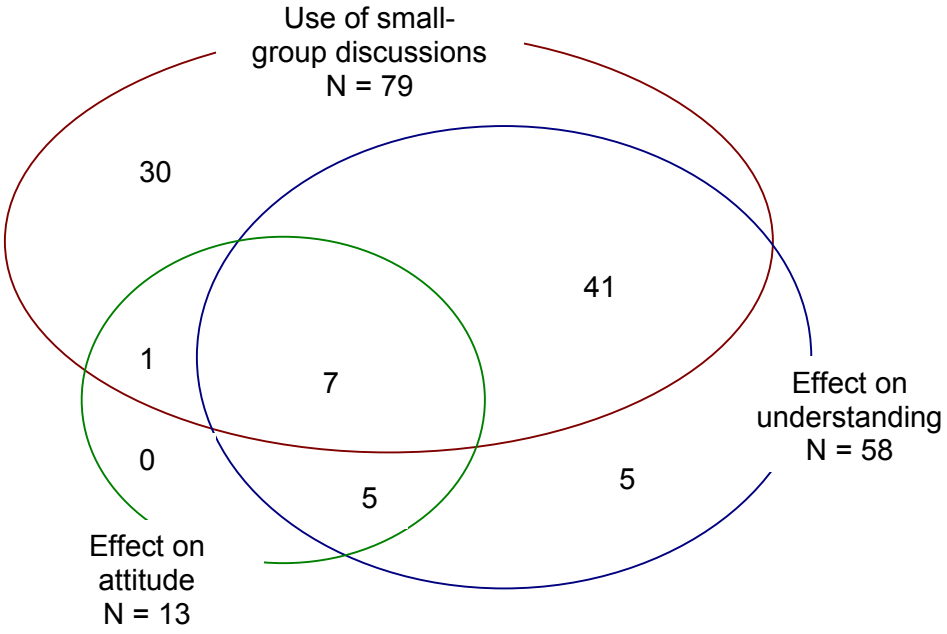
Table 3.2: Publication date of studies included in the systematic map (N = 89, mutually exclusive)

Publication period	Number of studies	%
1980 – 1985	1	1
1986 – 1991	5	6
1992 – 1997	36	40
1998 – 2002	47	53
Total	***89***	***100***

Table 3.2 indicates that the research activity in the review area has been minimal up to ten years ago and has been 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.

3.2 Characteristics of the included studies (systematic map)

Figure 3.2: Focus of the included studies (N = 89)



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 89 studies included in the review. Not surprisingly, almost all studies (79) report on the process of small-group discussions, although only 30 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 58 studies report on students' understanding of science, with only five of these only 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 (seven) reporting on all three aspects of the review question.

Table 3.3: Country in which the study carried out (N = 89, not mutually exclusive)

Country	Number of studies	% of the 89 studies
USA	35	39
UK	12	13
Canada	11	12
Netherlands	4	4
Australia	5	4
Germany	4	4
Hong Kong	4	4
Taiwan	4	4

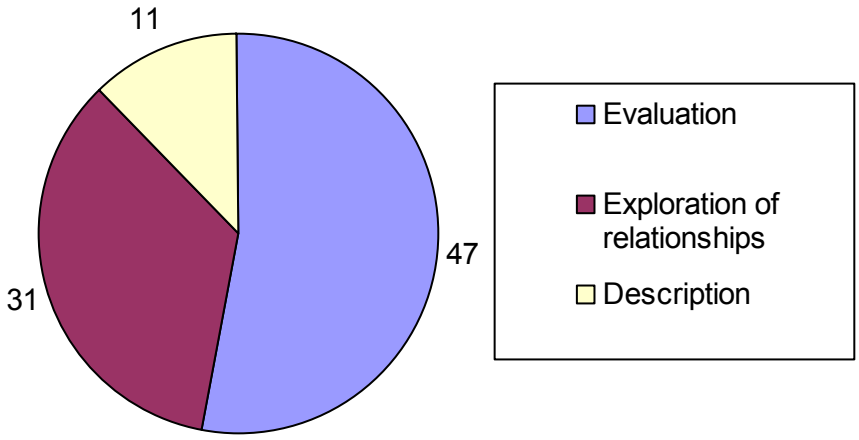
Country	Number of studies	% of the 89 studies
Spain	3	3
Finland	2	2
France	2	2
Israel	2	2
Greece	1	1
Malaysia	1	1
Brazil	1	1
Singapore	1	1
Total	***92***	

Data in Table 3.3 demonstrate that the 89 studies included in the map report on studies carried out in a large number of different countries. 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, Portugese 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 US 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) 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 (79) focus on secondary school learners between 11-16 of age; 18 also report on the age range 17-20. The learners in most samples (82) were of mixed sex. A total of three and 10 studies report on female- and male-only educational settings respectively.

Figure 3.3: Study type (N = 89)



EPPI uses a hierarchical 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 (47) of the studies report on evaluation studies, split almost equally between naturally-occurring (22) 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 (31) present explorations of relationships between different characteristics of small-group discussions. A minority of studies (11) provide only descriptions of small-group discussions.

Table 3.4: Distribution by discipline (N = 89, not mutually exclusive)

Science subject	Number of studies	% of the 89 studies
(Integrated) science	36	40
Biology	18	20
Chemistry	4	4
Physics	33	37
Earth science	4	4
Total	***95***	

Table 3.4 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 3.5: What types of learners are involved? (N = 89, not mutually exclusive)

Ability of learners	Number of studies	% of the 89 studies
Mixed ability	77	87
Lower ability/slow learners	4	4
Upper ability/gifted	11	12
Disaffected	2	2
Unspecified	1	1
Total	***95***	

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 mixibility levels in all but one case.

Table 3.5 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 3.6: Number of students in a discussion group (N = 89, not mutually exclusive)

Group size	Number of studies	% of the 89 studies
Groups of 2 (dyads)	28	31
Groups of 3 - 4	56	63
Groups of 5 - 6	13	15
Unspecified	10	10
Total	***107***	

Table 3.6 indicates that one-third of the studies focus on groups of two students, and 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 3.7: How discussion groups are constituted (N = 89, not mutually exclusive)

Group composition method	Number of studies	% of the 89 studies
Friendship ties	14	16
Randomly, by teacher	11	12
Randomly, but same sex groups	6	7
Purposely same ability	5	6
Purposely heterogeneously	29	33
Unspecified	37	42
Total	***102***	

The way discussion groups have been constituted is given in Table 3.7.

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) when pupils 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, pupils were allowed to form friendship groups as this is considered to encourage discussion (Gayford, 1995, p 136).

About 16% of all studies (14) allow groups to emerge from friendship ties and 42% of all studies (37) 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, fewer than half 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 pupils 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 pupils 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 (84) concern studies of self-contained and permanent groups, with only three studies considering snowballing and two jigsawing as an organisational structure.

Table 3.8: Duration of the discussion tasks (N = 89, not mutually exclusive)

Duration of the discussion tasks	Number of studies	% of the 89 studies
2 - 5 minutes	1	1
6 - 30 minutes	8	9
Close to class period (30 - 60 minutes)	30	34
Longer than a class period	30	34
Unspecified	24	27
Total	***93***	

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 3.8 shows that two-thirds of the studies (60) 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 3.9: Stimulus for discussion tasks (N = 89, not mutually exclusive)

Nature of the stimulus for the discussion task	Number of studies	% of the 89 studies
One-line oral teacher instruction	2	2
Oral context provided by teacher only	3	3
Newspaper article	1	1
Prepared curriculum print materials	59	66
Practical work	37	42
Computer software	22	25
Field trip	1	1
Video/TV/film clip	8	9
Learner generated	14	16
Total	***147***	

Discussion tasks in half of the studies used more than one type of stimulus. In two-thirds of the studies (59), 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 a quarter of the studies (22) 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 chat rooms 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 3.10: Product of discussion tasks (N = 89, not mutually exclusive)

Product of the discussion task	Number of studies	% of the 89 studies
Individual sense-making	83	93
Report group views/presentation orally in class	20	22
Support a group position in a class debate/quiz	10	11
Present group written project (including poster)	11	12
Other	6	7
Total	***130***	

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. In just under half of the cases (41), 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 3.11: What outcomes are reported (N = 89, not mutually exclusive)

Reported outcome	Number of studies	% of the 89 studies
Conceptual understanding of science	65	73
Evidence (methods and nature of science)	34	38
Applications of science	3	3
Attitudes to (school) science	14	16
Skills (communication/collaboration)	55	62
Decision-making on socio-scientific issues	10	11
Total	***181***	

As can be seen in Table 3.11, 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 two-fifths (38%) were interested in the understanding of evidence. Not surprisingly over a half of the studies (62%) 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 3.12: Number of discussion groups included in the study (N = 89, not mutually exclusive)

Number of groups in the study	Number of studies	% of the 89 studies
1 discussion group only	8	9
2 discussion groups	5	6
3 - 10 discussion groups	36	40
11 - 30 discussion groups	25	28
More than 30 discussion groups	15	17
Unspecified	5	6
Total	***94***	

The majority of studies involved three or more discussion groups with the highest number (36) being in the range 3-10. 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 (eight) 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 by Roth and colleagues.

Table 3.13: Research strategy used (N = 89, not mutually exclusive)

Research strategy	Number of studies	% of the 89 studies
Experiment	28	31
Survey	15	17
Case study	48	54
Action research	3	3
Ethnography	3	3
Total	***97***	

It is surprising that more than half of the studies (48) in the review can be characterised as case studies. 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 3.14: Nature of the data collected (N = 89, not mutually exclusive)

Nature of the data collected	Number of studies	% of the 89 studies
Test results	40	45
External examination results	1	1
Written reports/questionnaires	32	36
Concept webs	5	6
(Dis)agreement scores (e.g. VOSTS)	3	3
Self-reports (diaries, interviews)	30	34
Group discussions (audiotaped)	40	45
Presentations	1	1
Observed behaviour (including videotaped)	59	66
Computer logs	14	16
Total	***225***	

On average, the studies present findings based on more than two different types of data. Half of the research reports on small-group discussions are based on audiotaped (40 studies) and/or videotaped (59 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. One-third of the studies used questionnaires (32) and interviews (30) for collecting data, respectively.

Table 3.15: Relationships between discussion stimulus and reported learning outcomes (N = 89, not mutually exclusive)

Nature of the stimulus for the discussion task	Reported learning outcome						Total
	Concepts	Evidence	Application	Attitude	Communi-cation skills	Decision-making skills	
One line oral teacher instruction	1	2	1	1	1	1	2
Oral context provided by teacher	3	1	-	-	-	-	3
Newspaper article	1	1	1	1	-	1	1
Prepared curriculum print materials	46	19	2	8	37	8	59
Practical work	22	16	-	5	25	-	37
Computer software	15	10	-	2	14	-	22
Field trip	1	1	-	1	1	-	1
Video/TV/film clip	7	3	-	4	6	1	8
Learner generated	11	8	-	2	9	-	14
Total	***87***	***61***	***4***	***24***	***93***	***11***	***147***

The cross-tabulation in Table 3.15 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 89 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.

Before the start of the search, 20 'seminal papers' (SPs) were identified by the Review Team with the view that any quality search should at least identify these SPs. The validity of the search strategy was confirmed when the search strings for

ERIC and BEI together produced 17 of these papers (seven in both ERIC and BEI). The remaining three SPs comprised books and reports. These three publications were included in the search through the list of 'papers through contacts'.

All 1,814 papers resulting from the electronic search were screened on title and abstract between three team members. The reliability of the screening was established by independent screening of a random 2.5% sample (45 papers) by the three team members and an EPPI-Centre member. The inter-screener reliability as measured by the frequency method and the Cohen's Kappa method is shown in Table 3.16. The Cohen's Kappa method has the advantage of compensating for chance agreement.

Table 3.16: Inter-screener agreement (include-exclude) for first screening (N = 45 papers)

	Frequency method		**Cohen's method**	
	Identical decisions	Inter-screener agreement	Cohen's Kappa coefficient	Inter-screener agreement
Screener 1-Screener 2	37	82%	0.333	Fair
Screener 1-Screener 3	35	78%	0.169	Poor
Screener 2-Screener 3	37	82%	0.444	Moderate
Screener 1-EPPI member	27	60%	0.091	Poor
Screener 2-EPPI member	33	73%	0.411	Moderate
Screener 3-EPPI member	33	73%	0.411	Moderate

Percentage inter-screener agreement is at an acceptable level (60%-82%), but Cohen's Kappa values for inter-screener agreement seem reasonable for all pairs, apart from those involving screener 1. As a result all 600 papers initially screened by screener 1 have been re-screened by screeners 2 and 3. All discrepancies between decisions of screeners 2, 3 and the EPPI team member were discussed and resolved.

After arrival of the hard copies, all 363 papers were screened independently by two of the team members using the same exclusion criteria. Only in 25 cases (7%) did disagreement on inclusion emerge. These cases were resolved after discussion.

One hundred and fourteen papers (89 studies) were keyworded, of which 23 papers (over 20%) were keyworded by at least two team members. An EPPI-Centre member also keyworded nine of these. Table 3.17 shows the inter-rater agreement for the keywording process.

Table 3.17: Inter-rater agreement for keywording (N = nine papers)

	Keywords on core keywording sheet	**Keywords on review-specific keywording sheet**
	Inter-rater agreement (%)	Inter-rater agreement (%)
Team member 2 and team member 3	90	87
Team member 2 and EPPI member	81	47
Team member 3 and EPPI member	84	46

Table 3.17 shows that agreement of keyword allocation between team members is consistently high for both the general and the review-specific keywords (90% and 87% respectively). The agreement is lower between the respective team members and the EPPI-Centre member, an expert in a field other than science education. This difference is particularly striking for the agreement on the review-specific keywords. The agreement of less than 50% for the review-specific keywords seems due to the difference in familiarity with science education learning outcomes, curriculum initiatives, task stimuli, and the possible organisation and products of discussion tasks.

4. IN-DEPTH REVIEW: RESULTS

4.1 Selecting the studies for the in-depth review

The systematic map, which was based on 89 studies, indicated that interest in researching small-group discussions arises from at least five different, partly overlapping, research areas:

1. *Pedagogy:* Interest in the use of group work as a strategy for teaching and learning within the movement of student-centred learning (co-operative and collaborative learning). The focus has shifted from the large body of (mainly US) research on classroom management aspect of the groups to the issue of what actually happens when students interact within these groups.
2. *Curriculum materials*: Interest in student interaction with, and learning effectiveness of, specific types of curriculum materials, especially ICT, practical work, projects, fieldwork.
3. *Curriculum content:* Interest in specific aspects of science understanding related to discourse, especially argumentation, and decision-making on socio-scientific issues.
4. *Social construction of knowledge:* Interest in documenting critical incidences in learning paths arising from the interactions of group members.
5. *Constructivist research:* Interest in exploring students' ideas and cognitive processes by making them voice their thinking.

Five broad areas appeared worthy of more detailed exploration in the in-depth review:

A. The nature of the stimulus provided for the group and its effect on the development of understanding
B. The use of small-group discussions in relation to the development of understanding of socio-scientific issues
C. Aspects to do with group composition, looking out, for example, for relationships between group size or gender balance within groups and development of conceptual understanding
D. The effectiveness of small-group discussions for different learning outcomes: for example, argument, decision-making
E. The use of ICT in small-group discussions

Emerging from the map, a total of 14 potential in-depth review topics, from across areas A-E above, were presented to the full meeting of the Review Group on 1 October 2003. There was overwhelming consensus that the highest priority should be given to four in-depth review questions, paramount amongst them:
What is the evidence from evaluative studies of the effect of small-group discussions on students' understanding of evidence in science (area D)?

A focus on small-group discussions, intending to improve understanding of evidence, was justified as the review end-users (that is, students) have serious problems

getting to grips with the concepts involved. A second set of end-users (that is, teachers and curriculum developers) will benefit from such a review focus when structuring and monitoring effective learning experiences.

The in-depth review question relates directly to two of the areas described in the conceptual framework in section 1.2, the development of scientific literacy, and ideas about evidence, as students' ability to draw on evidence to develop and support ideas is seen as one of the key components of scientific literacy.

The application of the exclusion criteria specified in section 2.3.1 resulted in 14 studies for the in-depth review.

Given the very considerable amount of time and effort which has been expended producing the systematic map, coupled with the increasing topicality of small-group discussions, it has been agreed that the second year of the review will focus on the two further inter-related in-depth review questions:

1. What is the nature of small-group discussions aimed at improving students' understanding of evidence in science (area D)?
2. When using different stimuli (that is, print materials, practical work, ICT, video/film), what is the effect of small-group discussions on students' understanding of evidence in science? (area A)

This first in-depth review has addressed the question of how effective small-group discussions are in improving students' understanding of evidence in science. The subsequent review will link patterns of success (and lack of it) to the nature of the small-group discussions concerned. The subsequent reviews will look into the contribution different types of stimuli make to students' learning of evidence.

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

Studies yielded

Application of the criteria described in section 4.1 yielded 14 studies for the in-depth review as follows:

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

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

Four 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:

- Tolmie and Howe (1993) and *Howe, Tolmie and Anderson (1991)
- 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 bibliography in 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.

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 US, with others as detailed below. The proportion of studies undertaken in the US is considerably larger than the proportion of studies from the US in the map.

Table 4.1: Countries in which the studies selected for in-depth review were carried out (N = 14, mutually exclusive)

Country	Number of studies	Study
USA	8	Finkel, 1996 Hogan, 1999b Keys, 1997 Lavoie, 1999 Palincsar *et al.*, 1993 Sherman and Klein, 1995a Suthers and Weiner, 1995 Williams, 1995
UK	2	Gayford, 1995 Tolmie and Howe, 1993
Canada	1	Lajoie *et al.*, 2001
France	1	De Vries *et al.*, 2002
Hong Kong	1	Tao, 2001
Israel	1	Zohar and Nemet, 2002

The researchers

Of the 14 studies, half appeared to be undertaken by single researchers, of whom one (Williams) was clearly identified as a practitioner researcher, three seemingly resulted from PhD studies (Finkel, 1996; Hogan, 1999b; Keys, 1997) and three were completed by post-doctoral or senior researchers (Gayford, 1995; Lavoie, 1999; Tao, 2001). Four of the studies were undertaken by pairs of researchers (Sherman and Klein, 1995a; Suthers and Weiner, 1995; Zohar and Nemet, 2002; Tolmie and Howe, 1993) and three by teams of three or four researchers (De Vries *et al.*, 2002; 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's *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, five studies were externally funded: De Vries *et al.* (2002), Lajoie *et al.* (2001), Palincsar *et al.* (1993), Suthers and Weiner (1995) and Tolmie and Howe (1993).

Subject focus

Half of the 14 studies in the in-depth review focused on small-group discussions in Integrated Science lessons, four in Biology, three in Physics and none in Chemistry. This constitutes a slightly higher proportion of Biology and lower proportion of Physics lessons in comparison with the 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, e.g. genetic engineering or Human Immunodeficiency Virus (HIV)-Acquired Immuno-Deficiency Syndrome (AIDS).

While all the studies involved students in using evidence, they were based on a range of science topics and different aspects of using evidence. Three studies addressed specific areas where students encounter difficulties in understanding science ideas (Finkel, 1996, genetics; Palincsar *et al.*, 1993, kinetic theory; Tao, 2001, mechanics). Two had a specific focus on socio-scientific issues (Gayford, 1995, the greenhouse effect; Zohar and Nemet, 2002, genetic engineering). Three involved predictions based on evidence presented (De Vries *et al.*, 2002, sound; Lavoie, 1999, biology; Tolmie and Howe, 1993, mechanics). Four looked primarily at scientific method (Hogan, 1999b, building theories and models from primary evidence on the nature of matter; Lajoie, 2001, confirming or refuting hypotheses on disease diagnosis; Sherman and Klein, 1995a, designing controlled experiments; Williams, 1995, model building using biological material). Reasoning and argumentation skills were of interest to Keys (1997; investigated the use of scientific reasoning skills in collaborative report-writing) and Suthers and Weiner (1995; developing scientific argumentation and reasoning using HIV-AIDS as a case study).

Ages of learners in studies

The studies were undertaken with a diversity of age ranges of learners, as summarised in Table 4.2 below. The ratio of studies between junior secondary (ages 11-15) and senior secondary level (ages 16-18) mirror that of all studies in the map.

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

Age range	Number of studies	Study
16 – 18	3	De Vries *et al.*, 2002 Finkel, 1996 Tao, 2001
13 – 15	9	Gayford, 1995 Hogan, 1999b Keys, 1997 Lajoie *et al.*, 2001 Lavoie, 1999 Sherman and Klein, 1995a Suthers and Weiner, 1995 Tolmie and Howe, 1993 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 their researches but a few varied group size at different stages. Half of the studies involved groups made up only of pairs of students (De Vries *et al.*, 2002; Keys, 1997; Lajoie *et al.*, 2001; Tao, 2001; Sherman and Klein, 1995a; Suthers and Weiner, 1995; Tolmie and Howe, 1993). Six studies used only groups of three or four (Finkel, 1996; Gayford, 1995; Palincsar *et al.*, 1993; Hogan, 1999b; Williams, 1995) and one study 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 give details of group size (Lavoie, 1999). It seems

that the studies involving smaller group sizes are over-represented in this set of studies, possibly because the development of understanding of evidence involves pitching views of participants against one another, which possibly can be done more effectively in smaller groups.

Grouping strategy

How groups were formed varied somewhat depending on the focus of the study. In some cases (Finkel, 1996; Gayford, 1995; Lajoie, 2001), friendship groups were preferred. As explained by Gayford, they can help promote some discussions, such as those involving socio-economic aspects of science. However, most researchers deliberately created heterogeneous groups (De Vries *et al.*, 2002; Hogan, 1999b; Keys, 1997; Palincsar *et al.*, 1993; Sherman and Klein, 1995a; Tao, 2001; Tolmie and Howe, 1993). In some cases, particular care was taken to promote argumentation by pairing pupils with differing abilities (Sherman and Klein, 1995a), level of understanding (Tolmie and Howe, 1993) or 'mental models' (De Vries *et al.*, 2002).

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

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

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

4.4 Synthesis of evidence

Approach to synthesis

This section synthesises the data-extracted from the 14 studies. Section 4.4.1 provides an overview of the aims of the studies. In section 4.4.2, methodological considerations are synthesised in order to permit judgements to be reached about the quality of the studies (weight of evidence A).

Section 4.4.3 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 4.4.4 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.

4.4.1 Overview of the studies

Aims of studies

Two particular features of the reports were apparent when considering the aims of the studies. Firstly, a characteristic of a number of the studies was that they had a diversity of aims, not all of which related to students' use of evidence in small-group discussion work. For example, one of the aims of the Gayford (1995) study was to relate science learned in school with the needs of society. Secondly, 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 but one of the studies focused on evaluation of intervention programmes (nine of them named) which had as one of their aims the promotion of small-group discussion activities. The exception, the study by Tolmie and Howe (1993), had as its main aim the investigation of the impact of gender composition of discussion groups on the process of exchange of opinions between pupils engaged on a science task, and if any of such differences actually matter in terms of learning outcomes (Tolmie and Howe, 1993).

Of the intervention evaluations, four studies (De Vries *et al.*, 2002; Lajoie *et al.*, 2001; Sherman and Klein, 1995a; 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 were as follows:

- to determine the effect of three characteristics - the nature of the topic, the nature of the task and the role of technology - of a specific CLE (related to sound) on students' dialogue when dealing with evidence (De Vries *et al.*, 2002)
- 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)

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

- to document the effect of the intervention on the ways in which students construct and revise conceptual and strategic knowledge successfully as they solve complex genetic problems (Finkel, 1996)
- to evaluate the effect of discussion-based learning on understanding of an environmental issue and on students' ability to distinguish between evidence and opinion (Gayford, 1995)
- 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)

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

- to evaluate the effect of an intervention, stressing the metacognitive and group strategic aspects of knowledge co-construction on students' collaborative scientific reasoning skills and their conceptual understanding (Hogan, 1999b);
- 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).

4.4.2 Methodological considerations

Study designs

The study designs were equally divided between naturally-occurring evaluations (De Vries *et al.*, 2002; Finkel, 1996; Keys, 1997; Palincsar *et al.*, 1993; Suthers and Weiner, 1995; Tao, 2001; Williams, 1995) and researcher-manipulated evaluations (Gayford, 1995; Hogan, 1999b; Lajoie *et al.*, 2001; Lavoie, 1999; Sherman and Klein, 1995a; Tolmie and Howe, 1993; Zohar and Nemet, 2002). Of the latter, only two (Hogan, 1999b; Sherman and Klein, 1995a) used a RCT design.

It should be noted that some 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. Similarly, Finkel's (1996) main focus was on the types of knowledge used by students when participating in small-group discussions, whilst in the process she evaluated, as part of tracking changes in the use of these types of knowledge, effective group discussion strategies.

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, Hogan, 1999b; Lajoie *et al.*, 2001; Tolmie and Howe, 1993; 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, apart from De Vries *et al.* (2002) who used volunteers. 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 De Vries *et al.* (2002), Finkel (1996), 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.

Five studies (De Vries *et al.*, 2002; Finkel, 1996; 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 (Lavoie, 1999; Williams, 1995), and half of the studies (Gayford, 1995; Hogan, 1999b; Lavoie, 1999; Palincsar *et al.*, 1993; Sherman and Klein, 1995a; Tolmie and Howe, 1993; Zohar and Nemet, 2002) used quite sizeable samples, involving eight to ten classes.

The majority of studies provided limited information about the characteristics of students in the sample. Some samples were clearly atypical: for instance, highly motivated students (De Vries *et al.*, 2002; Tao, 2001), mainly from lower socio-economic backgrounds (Lavoie, 1999; Suthers and Weiner, 1995) or from a private girls school (Lajoie *et al.*, 2001).

Hardly any of the studies explicitly stated to what extent the findings were thought to be generalisable. Only Keys (1997) specifically warned against generalising her findings from her interpretive study as the effects of the intervention under scrutiny, and Lavoie (1999) limited claims to classes taught by teachers trained for, and committed to, the specific intervention. However, it seems many studies made an implicit claim for generalisibility: for instance, by motivating the study on the grounds

of changes in the national curriculum (Tolmie and Howe, 1993) or as a response to research trends (Palincsar *et al.*, 1993).

Comparison/control of independent variable

Six sizeable studies (Gayford, 1995; Hogan, 1999b; Lavoie, 1999; Sherman and Klein, 1995a; Tolmie and Howe, 1993; Zohar and Nemet, 2002) and one small study (Lajoie *et al.*, 2001) included a comparison group. Half of these (Gayford, 1995; Lavoie, 1999; Zohar and Nemet, 2002) compared the learning effect for groups undergoing an intervention with small-group discussion work, against those who learn through a traditional learning sequence. Three 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. For instance, Hogan (1999b) studied the effect of a collaborative reasoning skills course, 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. Two studies compared groups with different sample characteristics, such as gender (Tolmie and Howe, 1993) or ability (Sherman and Klein, 1995a), undergoing the same intervention.

Several of these studies carefully matched the experimental and control groups for teacher (Gayford, 1995; Hogan, 1999b; Lajoie *et al.*, 2001; Lavoie, 1999; Zohar and Nemet, 2002), students' prior conceptual understanding (Gayford, 1995; Lavoie, 1999; Hogan, 1999b; Zohar and Nemet, 2002), their science achievement (Gayford, 1995; Hogan, 1999b), gender (Hogan, 1999b), and subject preference (Gayford, 1995).

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. Tolmie and Howe's (1993) study used discussion groups consisting of all-male, all-female or mixed pairs. Each of the clusters in the latter study was controlled for the range of ability differences calculated as the 'coefficient of dissimilarity' of conceptual understanding between the pair.

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

Pre-post data-collection of dependent variable

Eight 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), Lavoie (1999) their logical thinking skills, and Tolmie and Howe (1993) their explanatory skills. Several studies measured effect by pre-post testing other variables. For instance, pre-post intervention tests (Gayford, 1995; Lavoie, 1999; Palincsar *et al.*, 1993; Zohar and Nemet, 2002) and pre-post intervention interviews (Keys, 1997) were used for documenting changes in conceptual understanding.

Some studies seemed to document benchmark data, but these were not comparable with the outcome data. For instance, Tao (2001) reported pre-intervention problem-solving success and post-intervention levels of conceptual understanding. More

subtly, Hogan (1999b) measured domain-specific knowledge before her intervention, and the ability to apply domain-specific knowledge after her intervention, where the latter may, at most, be part of the former.

Sherman and Klein (1995a) and De Vries *et al.* (2002) 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.

Four studies do not report benchmark data to be compared with the outcomes (Finkel, 1996; Lajoie *et al.*, 2001; Suthers and Weiner, 1995; Williams, 1995).

Reliability and validity of data-collection methods and tools

Only one 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.

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).

Several studies used tools with multiple items measuring the same concept, thus implicitly increasing the reliability (Finkel, 1996; Tao, 2001; Tolmie and Howe, 1993) but no inter-item reliability score was 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'.

Hogan (1999b) designed a pre-test (Perspective on Learning Science (POLS)) and through one-way analysis of variance (ANOVA) established that the results were independent from general science achievement ($F(1,161) = 1.50$, $p = 0.22$).

It seems that piloting data-collection instruments and strategies probably occurs more often than it is reported. Only two studies reported field-testing the instrument, and three the procedure. Gayford (1995) piloted his written tasks for comprehensibility and level of difficulty with an equivalent student group outside his sample. Tolmie and Howe (1993) used prediction tasks already tested in a previous study, thus serving as a pilot validity check. De Vries *et al.* (2002), Finkel (1996) and Keys (1997) provided an exercise for students to practise computer-based data-collection, group discussions and collaborative report-writing respectively.

Some research designs in themselves provide validity. For instance, the action research design of Suthers and Weiner (1995) uses tasks which are modified and

sharpened up for each subsequent research cycle, thus increasing the validity of the data over the lifetime of the project. Also, 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.

A number of studies provided little or no detail for judging the reliability or validity of the data-collection method and tools (De Vries *et al.*, 2002; Palincsar *et al.*, 1993; Williams, 1995).

Reliability and validity of data analysis methods

All but three of the studies (Finkel, 1996; 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.

Three studies (De Vries *et al.*, 2002; Palincsar *et al.*, 1993; Lavoie, 1999) used t-tests for identifying the significance of differences in the type of dialogue between two phases of an intervention (De Vries *et al.*, 2002), 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). All three studies provided details of group sizes, mean scores and standard deviations, t-values and p-values. For instance, Palincsar *et al.* (1993) 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.

Five studies (Lajoie, 2001; Hogan, 1999b; Sherman and Klein, 1995a; Tolmie and Howe, 1993; 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 most of these studies (Lajoie, 2001; Hogan, 1999b; Sherman and Klein, 1995a) provided details of group sizes, mean scores and standard deviations, F-values, degrees of freedom and p-values. Tolmie and Howe (1993) and Zohar and Nemet (2002) lack some information on standard deviations and/or F-values which makes it difficult to verify the conclusions drawn. For example, using ANOVA, Hogan (1999b) established a significant difference in post-intervention performance of the experimental over the control group ($F(3,159 = 4.02$, $p = 0.05$)). The conclusion that this effect is stronger for the learner-as-explorer than for the learner-as-student is also statistically supported. Sherman and Klein (1995a) 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.

The Mann-Whitney two-sample test was used by Gayford (1995) for comparison of pre-post performance for experimental and control groups for questions on the greenhouse effect and on the contribution of radiation. Statistically significant differences emerged in favour of the experimental group for middle and high ability learners (mean and standard deviations are provided). The same test was seemingly used for motivation scores (only means provided) with the same outcome.

Gayford (1995) reported that the Mann-Whitney test shows no statistical difference for students' ability to distinguish between opinion and scientific evidence between the experimental and control groups. However, no data were presented.

Sherman and Klein (1995a) 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) and Tolmie and Howe (1993) were based on Pearson correlation analysis. For instance, Tolmie and Howe (1993) reported a significant correlation between coefficients of dissimilarity (the difference in initial conceptual understanding of the group members) on the one hand, and the change in explanatory understanding ($r=0.19$, $p=0.05$) and the number of references to explanatory factors ($r=0.29$, $p=0.05$) on the other. Tolmie and Howe (1993) used Blalock's method of 'causal analysis' of the correlation of behaviour at different stages of a group discussion.

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. The study by Tolmie and Howe (1993) used the same proportion of written scripts and reported an agreement level of 90%. Zohar and Nemet (2002) reported at least 85% inter-rater agreements for an unspecified percentage responses to the three tests they used.

Tolmie and Howe (1993) provided an example of a reliability check for observation data. They describe independent coding of 25% of student-interaction transcripts with an 81% initial inter-judge agreement. Similarly, Keys (1997) uses 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), Tao (2001), and Tolmie and Howe (1993) 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 studies by Finkel (1996) and Keys (1997) described triangulation for validation of assertions in detail and to great effect.

Grounded theory has been used in five studies for developing categories of interactions and use of knowledge during group discourse (Finkel, 1996; Keys, 1997; Lajoie, 2001; Palincsar *et al.*, 1993; Tolmie and Howe, 1993). Keys (1997) mentions that she used Kuhn's framework of reasoning strategies as a basis for her grounded theory analysis of clinical interviews, thus increasing the validity. Similarly, Finkel (1996) used Perkins/Simmons knowledge frames as basis for analysing her data, and Lajoie (2001) used data from 'experts' for determining his typology for student performance in scientific reasoning.

An interpretive study like that by Keys (1997) will not focus on the reliability of the analysis of the data, as the intention is to provide as full a picture as possible, crystalising the information around a limited number of assertions supported by descriptive data. Finkel (1996) used the same method equally well.

Apart from a description of the statistical methods, several studies provided little or no detail of issues related to reliability or validity of the data analysis (De Vries *et al.*, 2002; Gayford, 1995; Lajoie *et al.*, 2001; Palincsar *et al.*, 1993; Sherman and Klein, 1995a; Suthers and Weiner, 1995; Williams, 1995).

Weighted evidence

Taking account of the different methodological aspects above, the quality of the 14 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 exercise, namely *'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)
De Vries *et al.*, 2002	Medium
Finkel, 1996	Medium-high
Gayford, 1995	Medium-high
Hogan, 1999b	Medium
Keys, 1997	Medium-high
Lajoie *et al.*, 2001	Medium
Lavoie, 1999	Medium
Palincsar *et al.*, 1993	Medium
Sherman and Klein, 1995a	High
Suthers and Weiner, 1995	Medium-low
Tao, 2001	Medium
Tolmie and Howe, 1993	Medium-high
Williams, 1995	Low
Zohar and Nemet, 2002	Medium

4.4.3 Appropriateness of the studies' research design for the in-depth review (category B)

This section of the report synthesises the evidence from the 14 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 effects of 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' (1997) descriptive study, and this evaluation focuses 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) and Lavoie (1999) 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 seven largest studies (Gayford, 1995; Hogan, 1999b; Lavoie, 1999; Palincsar *et al.*, 1993; Sherman and Klein, 1995a; Tolmie and Howe, 1993; 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 five studies (De Vries *et al.*, 2002; Finkel, 1996; 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 De Vries *et al.* (2002), Lajoie *et al.* (2001), Lavoie (1999), Tao (2001), 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 six studies (Gayford, 1995; Hogan, 1999b; Lavoie, 1999; Sherman and Klein, 1995a; Tolmie and Howe, 1993; Zohar and Nemet, 2002) use a control group. Several of these studies carefully matched the experimental and control groups for a teacher

effect and for students' prior conceptual understanding (Gayford, 1995; Hogan, 1999b; Lavoie, 1999; Zohar and Nemet, 2002). The other two studies took great care in the control of external factors when constituting their small-groups according to prescribed characteristics (Sherman and Klein, 1995a; Tolmie and Howe, 1993).

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 Hogan (1999b), Lavoie (1999), Tolmie and Howe (1993), Zohar and Nemet (2002). Other studies used pre-post intervention measures to establish change in students' conceptual understanding or did not collect any benchmark data.

Appropriateness of addressing issues of reliability and validity in data-collection

Section 4.4.2 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 (1995a).

The check by an external expert of the content validity of the instruments for measuring the understanding of evidence used by Lavoie (1999) and Zohar and Nemet (2002) is worth mentioning. Five studies (De Vries *et al.*, 2002; Finkel, 1996; Gayford, 1995; Keys, 1997; Tolmie and Howe, 1993) 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 4.4.2 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 effect on students' conceptual understanding or their attitudes. Some t-tests (De Vries *et al.*, 2002; Lavoie, 1999), several of the ANOVA methods (Hogan, 1999b; Sherman and Klein, 1995a; Tolmie and Howe, 1993; Zohar and Nemet, 2002) and some correlation studies (Tolmie and Howe, 1993) are focused on students' understanding of evidence. Gayford's (1995) extensive statistics, on the other hand, focus mainly on conceptual understanding and the findings related to students' ability to differentiate between opinion and scientific evidence is supported by minimal data only.

Weighted evidence

Taking account of the different methodological aspects above, the quality of the 14 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)
De Vries *et al.*, 2002	Medium-low
Finkel, 1996	Medium-low
Gayford, 1995	High
Hogan, 1999	Medium-high
Keys, 1995	Low
Lajoie *et al.*, 2001	Low
Lavoie, 1999	Medium-low
Palincsar *et al.*, 1993	Medium-low
Sherman and Klein, 1995a	Medium
Suthers and Weiner, 1995	Low
Tao, 2001	Low
Tolmie and Howe, 1993	Medium-high
Williams, 1995	Low
Zohar and Nemet, 2002	Medium

4.4.4 Relevance of the studies' focus for the in-depth review (category C)

Further features of the study designs are selected for their appropriateness for the in-depth review question. These five features are discussed below and will each contribute to the weight of the evidence for category B. Similarly, aspects of the way in which the variables are formulated and explicated are selected for 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 14 studies will be weighted according to five aspects: the nature and specificity of the independent variable (small-group discussion), the nature and breadth of the dependent variable (understanding evidence), and the representativeness of the research context.

Nature of the independent variable (i.e. small-group discussions)
Two types of small-group discussions can be identified in the studies. First, several small-group discussions are arranged around information about a science-based situation for which group members have explicit conflicting predictions or explanations. The discussion intends group members to deal with evidence from within the group. These small-group discussions will be called 'internal conflict small-group discussions'. Other small-group discussions are stimulated by information from learning materials conflicting with a prediction or explanation agreed within the group. Alternatively, learning materials may present two conflicting predictions, explanations or justifications. These small-group discussions will be called 'external conflict small-group discussions'.

Most studies in the in-depth review evaluate the effect of part or whole of the following sequence of learning experiences, based on the advancing, challenging and justifying of opinions (Tolmie and Howe, 1993, p 192):

1. familiarisation with science-based situation or problem
2. formulation of individual prediction or explanation in writing
3. construction of joint prediction/explanation through an internal conflict small-group discussion
4. collection/provision of observational data
5. modification of prediction/explanation to reconcile the data through an external conflict small-group discussion
6. production of an agreed record of prediction/explanation

Five studies (Keys, 1997; Lavoie, 1999; Hogan, 1999b; Palincsar *et al.*, 1993; Tolmie and Howe, 1993) include both internal and external conflict small-group discussions following the sequence above. For instance, Tolmie and Howe (1993) ask pairs to discuss their individual predictions of trajectories of a falling object and record a jointly agreed graph on screen. The software package then provides the actual trajectory and students are tasked to explain the differences. The 'observational data' in step 4 frequently emerged from class presentations and class practical work (Keys, 1997; Hogan, 1999b; Lavoie, 1999; Palincsar *et al.*, 1993). In general, the guidance for structuring both types of small-group discussions was minimal or unspecified.

Three studies (De Vries *et al.*, 2002; Gayford, 1995; Sherman and Klein, 1995a) use steps 1-3 only (an internal conflict small-group discussion), with structured guidelines for identifying differences in individuals' predictions and explanations. In the first study, in particular, differences between individuals' explanations and predictions are highlighted by a CBI and dyads are led through these differences in sequence for explicit discussing, explaining, verifying and information-searching. However, three studies (Finkel, 1996; Suthers and Weiner, 1995; Zohar and Nemet, 2002) use steps 4-6 only (an external conflict small-group discussion). Some, but less, structured guidelines are provided. For instance, Finkel (1996) establishes with her students explanatory genetics models at the start and then provides conflicting computer-generated data. She gives guidance in the form of an algorithm for modifying models in order to accommodate new data, which students soon vary or modify.

The study by Tao (2001) has no 'conflict' as the basis of the group discussion, since the discussion is about recognising own perceptions in multiple correct problem solutions. The role of group discussions in the studies by Lajoie *et al.* (2001) and Williams (1995) are unclear.

It is striking that studies including an internal conflict small-group discussion take great care to compose heterogeneous groups (Keys, 1997; Palincsar *et al.*, 1993; Tolmie and Howe, 1993; De Vries *et al.*, 2002; Hogan, 1999b; Sherman and Klein, 1995a), but studies involving an external conflict small-group discussion (Finkel, 1996; Gayford, 1995) rely on friendship groups.

Specificity of the independent variable (i.e. small-group discussions)

Only one study took small-group discussion as the explicit independent variable.

Lavoie (1999) explored the effects of an introductory teaching phase where students were asked to make individual predictions and discuss these in small-groups. The study compared the effects on students' process skills and logical thinking ability for those who are and are not taught through such an introductory phase. This aspect of Lavoie's design is highly appropriate for the review. Two more studies (Gayford, 1995; Zohar and Nemet, 2002) compare the learning effects of a teaching approach with a focus on small-group discussions, although small-group discussions are not the major component of these interventions. For instance, Zohar and Nemet (2002) explore the difference between learning of genetics from textbooks (without small-group discussions) and learning the same materials through a unit on genetics dilemmas (including extensive small-group discussions) supported by an input geared at developing argumentation skills.

Several studies evaluate learning effects by comparing small discussion groups of a specific composition. For instance, Tolmie and Howe (1993) report on the effects of differences in groups' gender composition, and Sherman and Klein (1995a) on those of groups' ability composition.

The majority of studies focus on the learning effect of a particular support provided for small-group discussions. This could be support in the form of features of software packages (De Vries *et al.*, 2002; Lajoie *et al.*, 2001; Sherman and Klein, 1995a; Suthers and Weiner, 1995), metacognitive discussion strategies (Hogan, 1999b; Zohar and Nemet, 2002), structured discussion algorithms (Finkel, 1996; Palincsar *et al.*, 1993; Tao, 2001) or collaborative writing tasks (Keys, 1997). The role of group discussions in Williams' (1995) study is merged with many other variables. For these studies, it is difficult to isolate the variable 'small-group discussions' from the evaluations.

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; secondly, it requires developing models or claims; and thirdly it allows drawing on data to justify models, claims or arguments.

Two of the studies (Gayford, 1995; Lavoie *et al.*, 1999) have a main focus on engagement with data. The first study looks at the effect of small-group discussions on students' ability to differentiate between opinions and scientific evidence in an essay on the greenhouse effect. Although, in practice, students may well have had to justify their views, argumentation was not the focus of the study. The second study looks at the effect of making sense of individual predictions in small-group discussions on conceptual understanding.

Five studies (Finkel, 1996; Hogan, 1999b; Keys, 1997; Palincsar *et al*., 1993; Tolmie and Howe, 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; Finkel (1996), Hogan (1999b) and Keys (1997) look at model reconstruction in the light of a variety of sources of information. For instance, Keys (1997) 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 a model; and to co-ordinate data in a coherent body to support a model. Tolmie and Howe (1993) focused on explanatory understanding. They develop 13 indicators or explanatory factors to describe the on-task interactions of students attempting to make sense of their predictions in the light of supplementary data.

Four studies (De Vries *et al.*, 2002; Lajoie, 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 De Vries *et al.* (2002), 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 (1995a), Tao (2001) and Williams (1995) is much more obscure. The first study measures scientific reasoning skills and process skills, whilst the second is interested in change in problem-solving skills.

The representativeness of the research context

Most studies collect data in intact classrooms (Finkel, 1996; Gayford, 1995; Hogan, 1999b; Keys, 1997; Lajoie *et al.*, 2001; Lavoie, 1999; Palincsar *et al.*, 1993; Tao, 2001; Tolmie and Howe, 1993; Williams, 1995; Zohar and Nemet, 2002). This research context will facilitate generalisation of the findings. Two types of situations could compromise the generalisibility of studies' findings: firstly, one study (De Vries *et al.*, 2002) uses volunteers; and secondly, several studies increase the researcher's control over interfering variables by the use of unnatural experimental situations. For instance, Suthers and Weiner (1995) use pairs of orally interacting students on different computers instead of one pair per computer; De Vries *et al.* (2002) observed pairs of students working in a special room one at a time; and Sherman and Klein (1995a) asked clusters of their dyads to work outside normal class in a special laboratory.

Weighted evidence

Taking account of the different aspects above, the quality of the 14 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 study's focus (weight of evidence C)

Study	Relevance of the focus of the study for the in-depth review (weight of evidence C)
De Vries *et al.*, 2002	Medium
Finkel, 1996	Medium-high
Gayford, 1995	High
Hogan, 1999b	Medium-high
Keys, 1997	Medium-low
Lajoie *et al.*, 2001	Medium

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

4.4.5 Overall weighting

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
De Vries *et al.*, 2002	M	ML	M	M
Finkel, 1996	MH	ML	MH	M
Gayford, 1995	MH	H	H	H
Hogan, 1999b	M	MH	MH	MH
Keys, 1997	MH	L	ML	ML
Lajoie *et al.*, 2001	M	L	M	ML
Lavoie, 1999	M	ML	L	ML
Palincsar *et al.*, 1993	M	ML	ML	ML
Sherman and Klein, 1995a	H	M	ML	M
Suthers and Weiner, 1995	ML	L	M	ML
Tao, 2001	M	L	ML	ML

Study	Weight of evidence A	Weight of evidence B	Weight of evidence C	Weight of evidence D
Tolmie and Howe, 1993	MH	MH	M	MH
Williams, 1995	L	L	L	L
Zohar and Nemet, 2002	M	M	M	M

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

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. Additionally, all four core team 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 have been modified slightly.

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 effects of small-group discussions on students' understanding of evidence in science?

5.1.2 Mapping of all included studies

Eighty-nine 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 report work that has taken place in the US, the UK and Canada.
- Small-group discussions are used with all ages of student in the secondary age range.
- The majority of work focuses on small-group discussions in relation to students' understanding.
- A diversity of measures is 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 involve groups of three to four students emerging from friendship ties, and have a duration of at least 30 minutes.
- Typical small-group discussions have individual sense-making as their main aim (as opposed to, for example, leading to a group presentation) and use prepared printed materials as the stimulus for discussion.
- The most common research strategy was that of case study.
- Twenty-eight studies had experimental designs, of which 12 were RCTs.
- The most popular techniques for gathering data are observation, videotapes and audiotapes of discussions, interviews, questionnaires and test results.

5.1.3 Nature of studies selected for the in-depth review

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

Table 5.1: Overall weights of evidence assigned to studies

Overall weight of evidence	Number of studies	Study
High	1	Gayford, 1995
Medium-high	2	Hogan, 1999b Tolmie and Howe, 1993
Medium	4	De Vries *et al.*, 2002 Finkel, 1996 Sherman and Klein, 1995a Zohar and Nemet, 2002
Medium-low	6	Keys, 1997 Lajoie *et al.*, 2001 Lavoie, 1999 Palincsar *et al.*, 1993 Suthers and Weiner, 1995 Tao, 2001
Low	1	Williams, 1995

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

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

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

(a) The use of small-group discussions based on a combination of internal conflict (that is, where a diversity of views and/or understanding are represented within a group) and external conflict (where an external stimulus presents a group with conflicting views) resulted in a significant improvement of students' understanding of evidence (from Tolmie and Howe, 1993).

(b) Improvement of students' understanding of evidence was not significantly different for members of all-female, all-male or mixed gender pairs. The benefit was greatest for female students when they were given several opportunities to engage with aspects of tasks related to understanding of evidence (from Tolmie and Howe, 1993).

(c) Improvement of students' understanding of evidence correlated with the initial *dissimilarity* of the group members in terms of their domain-specific understandings: that is, student groups were constructed in such a way that they contained students with as wide a range of domain-specific understandings as was possible (from Tolmie and Howe,1993, and supported by findings of De Vries *et al.*, 2002, who also constructed their pairs for maximum dissimilarity).

(d) The use of small-group discussions did not affect students' ability to differentiate observational or experimental data from opinions in a science-based text (from Gayford, 1995).

(e) The use of small-group discussions supported by a specific programme fostering collaborative reasoning (including evaluating and strengthening of knowledge claims) improved students' metacognitive knowledge of collaborative reasoning (including their knowledge of reasoning about evidence) significantly more than for students not following the special programme. However, such gain within the treatment group depended on learners' perspective on learning: students with a *learner-as-explorer* perspective gained significantly more than peers with a *learner-as-student* perspective (from Hogan, 1999b).

(f) The improved metacognitive knowledge of collaborative reasoning described above did not translate into better use of strategies while reasoning, including when dealing with scientific evidence (from Hogan, 1999b).

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

(g) The use of internal conflict small-group discussions (from De Vries *et al.*, 2002) *or* external conflict small-group discussions (from Finkel, 1996, and from Gayford, 1995) produced improvement in students' understanding of evidence.

(h) The use of small-group discussions (together with specific instruction in argumentation skills) improved students' ability to construct more complex arguments (from Zohar and Nemet, 2002).

(i) The effectiveness of small-group discussions in producing an improvement in students' understanding of evidence depended on three types of understanding: understanding of the science domain, the process by which model-revision takes place and metacognition (from Finkel, 1996).

(j) The use of small-group discussions resulted in a significantly higher achievement in understanding of evidence for students using a cued version (that is, one which gives students specific instructions on what to include in points they make in discussions) of a computer-based instruction (CBI) program than a non-cued version (from Sherman and Klein, 1995a, and from De Vries *et al.*, 2002). This evidence was strengthened by the more general findings of Finkel (1996) and Palincsar *et al.* (1993) that a scaffolding routine for structuring small-group discussions improved students' understanding of scientific evidence.

Beyond the specific focus of the in-depth review question, one additional finding worth noting is that there was reasonable evidence to suggest that the gender composition of small discussion groups determined the interaction style for developing students' explanatory understanding. All-male groups confronted differences in their individual predictions and explanations; all-female groups searched for common features of their predictions and explanations across tasks; and mixed groups secured progress through turn-taking (from Tolmie and Howe, 1993, and possibly explaining the finding by De Vries *et al.*, 2002, that engagement with individually different views was required in small-group discussion in order to

impact on explanatory understanding, which occurred in the all-male, but not the all-female pair they describe).

Links with other reviews

No other reviews of small-group discussions in science lessons, systematic or narrative, have been undertaken. It was therefore not possible to compare the findings of this review with those of other reviews.

5.2 Strengths and limitations of this systematic review

Strengths

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.

- 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 was a scarcity of studies that focused on small-group discussions as a discrete independent variable, which resulted in very little work emerging which related specifically to the in-depth review question. Of these studies, only about half use a comparison group for small-group discussion as a teaching strategy. As a result, only seven studies were judged to be of reasonable quality with respect to the review question; that is, only seven studies had an overall weight of evidence of medium or higher.

- Although the studies in the in-depth review shared a number of similar characteristics at the broad level, there were considerable differences at the detailed level. For example, 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 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 Team feel some concern about the number of low quality studies which had to be included in the in-depth review, as judgements of quality are not made until comparatively late in the review process. However, this is a function of the process itself, rather than this specific review.

5.3 Implications

The Review Team 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 small-group discussions adversely affect students' understanding of the nature of evidence. Therefore there is nothing to suggest that current policy (which is strongly advocating the use of small-group discussion work) should be changed. 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.

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 teachers should 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 nature of the stimulus provided for the group and its effect on the development of understanding
- 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. argument, decision-making)
- the use of ICT in small-group discussions

The Review Group will explore some of these areas in its next review.

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.

6. REFERENCES

6.1 Studies included in map and synthesis

The 89 studies included in the systematic map were reported in 114 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 C-Y, Mao S-L (1999a) Comparison of Taiwan science students' outcomes with inquiry-group versus traditional instruction. *Journal of Educational Research* **92:** 340-346.

Chang C-Y, Mao S-L (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 H-P, 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.

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:** 757-92.

*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.

*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 W-M (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, 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 W-M, 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 W-M, 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.

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 H-C (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 P-K (1999) Peer collaboration in solving qualitative physics problems: the role of collaborative talk. *Research in Science Education* **29:** 365-383.

Tao P-K (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 P-K (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 P-K, 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 C-C (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 JM, Hogarth SD, Lubben FE (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.* Issue 1. London: EPPI-Centre, Social Science Research Unit.

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.

Deparment 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.

Fensham PJ (1988) Approaches to the teaching of STS in science education. *International Journal of Science Education* **10:** 346-356.

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, Collins S (2001) Pupils' views of the role and value of the science curriculum*. International Journal of Science Education* **23:** 441-467.

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.

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 maths, technology, social science or computing*

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

3. Not about small-groups
- *2-6 participants*

4. Not on substantive and explicit discussion tasks
- *explicit discussion tasks taking more than 2 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 11-18, or main focus not on learners aged 11-18
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
Pupils aged 11 to 18

Limits
English language
1980 to date

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

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))

2.2.3 PsycINFO

PsycINFO was searched on 10 April 2003, using the WEBSPIRS interface and 537 records were retrieved.

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.

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 (1-10) 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 2.5: Indicators for weight of evidence

Review question:
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 M.11, 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 M.11, 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 ..... The study is an evaluation. If not, final weight for B: LOW If so, weighting according to aspects below			**For the RQs relevant to the review .....**			
sample size			**nature of sample**			Sum total and classification for D:
large sample with appr. sampling method	large sample, no sampling method	small sample (up to three classes)	highly representative of small group discussions	less representative of small group discussions	not representative of small group discussions	3-4: low 5-7: medium-low 8-10: medium 11-13: medium-high 14-15: high
comparison/control			**focus of intervention**			
comparison for SGD in design (control, types)	comparison for SGD in findings only	no comparison/control	SGD is sole&explicit independent variable	SGD is a major discrete element of intervention	SGD 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 rel/val for data-collection	some checks on rel/val for data-collection	little/no checks on rel/val for data-collection	reports broad range of understanding of evidence	reports narrow range of understanding of evidence	reports understanding evidence only indirectly	
data analysis:			**situation**			
solid checks on rel/val for data analysis	some checks on rel/val for data analysis	little/no checks on rel/val for data analysis	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 89 studies in the review according to the type of research study reported.

Table A lists the 11 reports of descriptive studies.

Table B provides an overview of the 31 studies reporting explorations of relationships.

Tables C and D list the reports of the 22 naturally-occurring and 25 researcher-manipulated evaluative 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 in 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 on reports of attitudinal aspects, communication skills of group members, and decision-making skills on socio-scientific issues is listed.

Table A: Summary of reports of descriptive studies included in the review (N = 11)

		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	✓			✓					
1183	Osborne *et al.*, 2001	✓				✓				✓
1322	Richmond and Striley, 1996	✓				✓				
481	Fawns and Salder, 1996	✓							✓	
977	Looi and Ang, 2000	✓							✓	
1377	Roth, 1999	✓							✓	
1398	Roychoudhury and Roth, 1996	✓							✓	

Table B: Summary of reports of studies exploring relationships included in the review (N = 31)

		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	✓				✓				
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	✓	✓		✓	✓			✓	

		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
1389	Roth and Roychoudhury, 1992	✓	✓		✓	✓			✓	
1544	Stein, 1997	✓	✓			✓				
693	Hogan, 1999a	✓	✓		✓	✓		✓	✓	
1632	Tsai, 1999	✓		✓		✓		✓	✓	
1824	Osborne *et al.*, 2002	✓	✓	✓	✓	✓	✓	✓		✓
1457	Seiler *et al.*, 2001	✓	✓	✓	✓	✓		✓	✓	
1514	Solomon, 1992	✓	✓	✓	✓			✓	✓	✓

Table C: Summary of reports of naturally-occurring evaluative studies included in the review (N = 22)

		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, 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		✓	✓	✓		✓	✓		
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	✓	✓	✓	✓	✓		✓		

		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
1471	Sherman and Klein, 1995a	✓	✓	✓		✓		✓	✓	
258	Chang and Lederman, 1994		✓		✓				✓	

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

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.	
Country of study	Not stated but assumed to be France
Details of researchers	Researchers were academics from two French universities funded in part by an EU grant.
Name of programme	CONNECT Confrontation, Negotiation, and Construction of Text
Age of learners	16 to 17
Type of study	Evaluation: naturally-occurring
Aims of study	To determine the factors that must be taken into account in designing a computer-supported collaborative learning situation that encourages students to discuss scientific notions. These include the nature of the topic (sound), the nature of the task (dealing with evidence by dialogue) and the role of technology (computer-supported learning).
Summary of study design, including details of sample	Intervention: Phase 1: Each student comments on responses to specific questions by both dyad members. Depending on the overlap of individual responses, dyads are asked to discuss, verify, explain or refer their responses. Phase 2: Dyads are requested to develop joint written responses to the questions. Discussion turns are logged and classified according to 13 categories within explanation, argumentation, problem-solving and management. Actual sample: 14 (out of 15 volunteers) were chosen to work in groups of two. In six cases, the pairs worked synchronously on the task but in different rooms. In the seventh case, the students worked synchronously side by side as a pilot.
Methods used to collect data	• Self-completion report or diary • For identifying student differences (Phase 0), students were asked to write an individual interpretation of a physical phenomenon that they had been given by text and figure (two-tambourine situation). • Data for intervention (phases 1 and 2) was collected by computer log.
Data-collection instruments, including details of checks on reliability and validity	• Task sheet regarding two-tambourine situation. • CONNECT sequences for phase 1: commenting on original text of both dyad partners and guided discussion of responses on specific questions; for phase 2: task for constructing joint text. Checks on reliability: None Checks on validity: Validity of data-collection was not explicitly discussed but was whole of actual dialogues of students working on their tasks. There is a pilot exercise the students go through, so they are familiar with the IT environment.
Methods used to analyse data, including details of checks on reliability and validity	• Classifying written predictions and identifying contrasting view for dyad composition • Identifying combinations of answers of specific questions for initiating guided discussion • Generating coding scheme of dialogue turns in 13 categories, with four main categories explanation, argumentation, problem-solving and management • Frequency counts/percentage of dialogue turns and task actions for both phases • Frequency counts/percentage of use of argumentation/ explanation/ management in both phases • Statistical significance of differences in occurrence of argumentation/ explanation/ management in both phases Statistical methods used: frequency counts, percentages; t-test for significance testing Statistical tests were applied to the quantitative data (Dialogue Turns and Task Actions) from six pairs

	Checks on reliability: Use of standard statistical test (t-test). For identifying student differences (Phase 0), the three researchers jointly rated all 15 texts. Phases 1 and 2 involved full record of student dialogue when discussing experiment and agreeing common texts. Checks on validity: three authors jointly analysed the whole corpus (a total of 492) collective discussions in six dialogues). Analysing whole of data collected from student dialogues
Summary of results	• Topic domain (sound) - Episodes in which the occurrence of epistemic dialogue was closely related to levels of description, different perspectives and double meanings in the domain and as such contributed to the development of conceptual understanding in that domain. • Task sequence - The task sequence procedure maximised the chances for students to have different conceptions and models. However ,putting students together with different viewpoints is not a sufficient condition. Students must notice their differences and want to discuss them. • The CONNECT interface helped students gain an understanding of their partner's views, reflect upon them and compare them with their own. The quantitative analysis of the interactions showed a prevalence of dialogue over task actions. This predominance was viewed as a positive outcome of the design of the interface and task sequences. Due to the burden of communication in a computer-mediated situation, task actions could well have prevailed over dialogue. • For some students, conceptual understanding can take place through conceptual differentiation resulting from the resolution of vocabulary ambiguities. For other students, dialogue leads to the recognition of a lack of understanding. For other students again, dialogue does not lead to understanding but is a missed opportunity.
Conclusions	• How different components of CSCL environments can play a role in favouring epistemic dialogue. • There is a complex and interacting set of factors that are involved in enabling students to engage in such dialogues in a way that could lead to conceptual understanding and have described way in which this can take place.' • CONNECT provides more focused development of explanation and argumentation than reported in similar studies with other software.
Weight of evidence A (trustworthiness in relation to study questions)	Medium The quality of data-collection and particularly data analysis is high. The small sample and the use of volunteers precludes generalisibility, and would possibly suggest reporting the effect of the different features of CONNECT for the different dyads, rather than across the dyads.
Weight of evidence B (appropriateness of research design and analysis)	Medium-low The sample size is small (14 students); the design does not include a control group and records the understanding of evidence longitudinally, but no bench data. Little information is provided on the reliability and validity of the data-collection method, although a pilot was used; high reliability and validity for the analysis.
Weight of evidence C (relevance of focus of study to review)	Medium The study uses ability-based discussion groups. It links the understanding of evidence to the features of the software package, not to group discussions. The outcome variable measures understanding of evidence directly and over some breadth (explanation and argumentation). The study relied on volunteers outside class and in different rooms.
Weight of evidence D (overall weight of evidence)	Medium

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.	
Country of study	USA
Details of researchers	PhD researcher at the University of Wisconsin-Madison
Name of programme	Not applicable
Age of learners	Not explicitly stated but likely to be 16 to 18
Type of study	Evaluation: naturally-occurring
Aims of study	To uncover ways in which students collaborate to construct, use and revise conceptual and strategic knowledge as they solve complex genetics problems
Summary of study design, including details of sample	Sequential data-collection from eight 'research groups' (three or four members each) in one class, each working on three or four tasks providing genetics data and from group presentations of revised models presented and critiqued Exam taken co-operatively at the end of the 1st phase gave students the opportunity to demonstrate their ability to use the models. Taped group discussions, computer logs, individual diaries and student work have been collected. Also plenary class presentations and discussions are tape-recorded. Actual sample: 25 students, in eight groups of three or four
Methods used to collect data	• Observation of: audio-recorded oral group interactions during model revision; audio-recorded whole class presentations and discussions • Data collected for measuring the variables: computer logs of actions during model revision; written materials produced during model revision
Data-collection instruments, including details of checks on reliability and validity	Instruments used: as above Checks on reliability: collecting discussion data from three tasks aiming at the same variables, provides reliability of the method; gathering data on the same event through different sources (discussions, logs, written reports) increases the reliability. Checks on validity: recording whole conversations, keeping computer records and student written work - all direct from the students Students had prior experience in Phase 1 of the method of recording conversations and so were comfortable with that.
Methods used to analyse data, including details of checks on reliability and validity	Grounded theory is used, in phase 1, resulting in: • indicators for the different variables (use of three types of knowledge); • set of 10 standard descriptors of the use of knowledge. These in turn were used as a framework in phase 2, resulting in: • narrative descriptions of each group's work on each of the tasks. Frequency counts for each group per task of: • recognition of anomalies • number of models generated • final model generated No statistical methods used Checks on reliability: Triangulation increased reliability. Checks on validity: One assumes the supervisor has been involved in the analysis, increasing the validity.
Summary of results	Three kinds of knowledge are used during model revision: • knowledge of genetics: for recognising anomalies in sets of data, and for the use of templates as starting point for model revision • knowledge of the process of model revision: guiding the way of revising models - derived from a set of ideas about the nature of science and the nature of models, which affected their view of how to revise a model, and secondly from comments made by the teacher

	• meta-cognitive knowledge of problem-solving strategies: for monitoring the revision process, and linking new models and their knowledge of genetics
Conclusions	Conclusions are similar to the findings, apart from the teaching implications below: • Students' emphasis on finding the right, final answer whereas the teacher was trying to emphasise that the focus of the activity was on process rather than product. • The type of genetic knowledge NOT used by students, in this case about meiosis. The role of the teacher is important in offering suggestions for tools and strategies. • Students rarely referred to models they had themselves created previously; they preferred Mendel's formal, clearly represented model rather than other less clearly and formally represented.
Weight of evidence A (trustworthiness in relation to study questions)	Medium-high The only drawbacks are the low generalisibility and the lack of information on how 10 descriptors were used, but the quality of the study is very good.
Weight of evidence B (appropriateness of research design and analysis)	Medium-low The study had a small sample of 25 students in one class; the design did not include a control group, and students' understanding of evidence was traced longitudinally but not pre-intervention; some measures were reported for reliability and validity of data-collection but hardly any for analysis.
Weight of evidence C (relevance of focus of study to review)	Medium-high The nature of the group discussions was representative but applied to an elective course; the independent variable of the study was group discussion, with very appropriate measures to document understanding of evidence in some breadth. The situation was highly representative of classroom learning.
Weight of evidence D (overall weight of evidence)	Medium

Gayford C (1995) Science education and sustainability: a case-study in discussion-based learning. *Research in Science and Technological Education* **13:** 135-145.	
Country of study	UK
Details of researchers	Researcher at the University of Reading
Name of programme	Not applicable
Age of learners	16
Type of study	Evaluation: researcher-manipulated
Aims of study	To evaluate the effect of discussion-based learning on understanding of an environment issue and on students' ability to distinguish between evidence and opinion
Summary of study design, including details of sample	Two classes were identified in each of four schools. One of each pair was the experimental class. Each class spent two 60-minute periods on the study (possibly more for some follow-up tests). Experimental classes were divided into groups of three or four students and encouraged to discuss the written material and tasks in the group. In a follow-up session 1, 2 or 3 days later, further individual and group

	tasks were done. Control classes worked as a whole class with similar materials and asked the teacher questions. Actual sample: No exact numbers given but eight classes of 21 - 27 students = about 192
Methods used to collect data	• Pre- and post-tests of six topic questions . • Self-completion questionnaire • Motivation measured on three-point scale • Written two-dimensional models of solar radiation scored on a 10-point scale • Not clear how data on aspect based on evidence or on opinion were recorded; how data on role of scientists was collected; how data on students' views on important measures for sustainable development were recoded
Data-collection instruments, including details of checks on reliability and validity	• List of questions used to test knowledge is given at the end of the paper along with comprehension text. Some ways of assessing outcomes are not described in detail. Checks on reliability: used pre-post test which was piloted on non-project students which looked at the appropriateness and comprehensibility of the materials used and some of the questions asked. Not for other measures Checks on validity: used test based on curriculum materials. Test and associated material provided were piloted for level of difficulty and application with students not involved in the project.
Methods used to analyse data, including details of checks on reliability and validity	• Pre-post test (a) was analysed by Mann Whitney U test. • Statistical significance levels are given for model making and motivation so assume same test but no data given. • Measures to control greenhouse effect discussed qualitatively with reference to higher reporting frequencies with experimental group but no data or analysis. • Aspect of statements made on basis of opinion or evidence but no data or statistics given. • Role of scientists described in qualitative way. Checks on reliability: no details Checks on validity: no details used standard statistical method; no details of qualitative tests
Summary of results	• Overall scores for pre- and but post-test questions showed significantly higher ($p > 0.05$) level of understanding at the end of the activity among those who were involved in the group activities compared with the control group. • Main differences ($p > 0.05$) were among the middle and lower groups; upper ability groups showed no statistical differences. • No statistical difference between experimental and control groups in identification of statements of opinion and of fact 'but it was notable this part of the activity did generate considerable discussion'. • Experimental group showed a statistically higher level of overall performance compared with the control group for construction of the two dimensional model. • Both groups showed consistent understanding of the important role of scientists in addressing the problems of the greenhouse effect. • Both groups showed a similar range of responses in relation to the measures that they felt would be necessary to control the greenhouse effect and would be sustainable but the reporting frequency was much higher in the experimental group. Only qualitative information, given actual frequencies not reported. • Motivation – The T & L activity adopted in the study was the most enjoyable and generally perceived as worthwhile by the experimental group and was statistically significant ($p > 0.05$) when compared with the control group.
Conclusions	• Students learned more effectively than a control group who worked individually. The gains were particularly marked among those of the middle and lower ability. A considerable amount of learning occurred in both types of group. • There was an appreciation of both the contribution and the limitations of science in addressing the phenomenon. • The majority of student s, particularly the lower 50% in terms of ability, performed significantly better in the experimental groups where

	discussion was encouraged. • The amount of questioning and answering that was possible in the experimental groups was far greater than would have been possible with a more traditional teacher-led session. • Motivation remained high throughout the activity. Motivation was also considerably greater among the experimental groups and this would have consequences for subsequent learning.
Weight of evidence A (trustworthiness in relation to study questions)	Medium-high Design is good for the main aspect reported on. More detail of school and pupils would also have helped. Six of the seven measures/aspects are reported in insufficient detail; this was deliberate as the authors were concentrating on (a) for this paper.
Weight of evidence B (appropriateness of research design and analysis)	High Good sample size taken from four schools; used control groups in paired classes; pre-post benchmarking data; pilot for data-collection for reliability and validity; no reliability or validity discussed for analysis but used standard tests.
Weight of evidence C (relevance of focus of study to review)	High The sample came from four schools and mixed gender; the prime focus was on focus is SGD and the measures were appropriate and sufficient. The intervention was carried out in the classroom and was representative of typical T & L.
Weight of evidence D (overall weight of evidence)	High

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.	
Country of study	Assumed USA
Details of researchers	Researcher at Institute of Ecosystems for component B. Teaching or other staff for components A and B.
Name of programme	Thinking Aloud Together
Age of learners	11 to 16
Type of study	Exploration of relationships Evaluation: researcher-manipulated
	To evaluate the effect of an intervention stressing the meta-cognitive and group strategic aspects of knowledge co-constructed on students' collaborative scientific reasoning skills and their conceptual understanding
Summary of study design, including details of sample	Mixed method design. Component A (quantitative): four intact equivalent treatment classes; four intact equivalent control classes; unit of measurement = individual outcomes. Controlled for school (same school), teacher (equal number of treatment/control classes from two teachers) and group composition (all heterogeneous for gender and ability) Component B (qualitative): purposively chosen four treatment and four control groups; unit of measurement = whole group performance. Checked on selection bias on prior equivalency variables, i.e. domain specific knowledge (nature of matter) with $F(1,144) = 0.73$, $p = 0.40$ and general science achievement with $F(1,161) = 0.18$,$p = 0.67$. Sample A: Actual sample of 163 students (81 treatment, 82 control). Sample B: subset of 24 observed in groups, subset of 12 in interviews.

Methods used to collect data	• One-to-one interviews and observation for B • Self-completion questionnaire: prior equivalency tests and MKA test • Psychological test: POLS • Hypothetical scenario including vignettes: APA test
Data-collection instruments, including details of checks on reliability and validity	Tools for prior equivalence variables (domain specific knowledge and general science achievement) are not specified. For component A: • POLS: seven written open response items, all specified. • APA: Part 1: individual written response to given problem-solving scenario. Part 2: discussion of individual responses with peer group. Part 3: individually revising/elaborating original response in Part 1. • MKA : Written responses to prompts related to six episodes of video of teenage actors collaboratively reasoning about a problem. (examples of prompts provided). For component B: • No tools were provided for the group tape/video recorded discussions. • No interview protocols were provided. Checks on reliability: Teachers followed a written protocol specifying method for the data-collection. Researcher observed several data-collection instances. Checks on validity: POLS: pilot run with previous year's cohort; discriminant validity check using current data for one-way analysis of variance of POLS versus general academic ability, concluding independence with $F(1,161)=1.50, p = 0.22$. APA: task adapted from Eishinger *et al*. (1991). No validity checks mentioned for prior equivalency variables instruments, for MKA tool or for qualitative collection instruments.
Methods used to analyse data, including details of checks on reliability and validity	Component A: • 2x2 ANOVA analysis of variance for POLS scores (per group) versus MKA scores. F max=2.96; ratio largest: smallest cell size=2.42, so homogeneity of variance also for POLS versus APA scores F max=1.92; ratio largest : smallest cell size=2.42, so homogeneity of variance. Component B: • Ethnographic micro-analysis of group interactions • Use of Erickson (1992) and Jordan and Henderson (1995) analysis schemes Checks on reliability: For component A: independent coding of 25% of all POLS and APA data by two researchers, Cohen's Kappa coefficient = 0.85 in both cases. Low inter-rater agreement on MKA coding (61%), so coding scheme re-validated (see below). For component B: No reliability measures reported for analysis of qualitative data. Checks on validity: Component A: validation of coding rubrics for MKA data between two researchers for 40 scripts. qualitative data triangulate quantitative findings. Component B: No validity measures reported for qualitative data.
Summary of results	• Students who received the intervention gained in meta-cognitive knowledge about collaborative reasoning and ability to articulate their collaborative reasoning processes compared to students in control classes. • This enhanced meta-cognitive awareness did not translate into improved collaborative reasoning behaviours, nor, therefore, into deeper processing of ideas and information that would have been manifest as enhanced ability to apply conceptual knowledge.
Conclusions	• Explicit teaching about collaborative scientific reasoning is required in order to help students articulate and evaluate their own and others' collaborative reasoning processes. • Students who view themselves as learner-as-explorer outperformed those with views of themselves as learner-as-student.

	• Treatment students do not use cognitive strategies any better in their reasoning, as evidenced on their conceptual understanding, in this case, of the nature of matter. Neither do they show a difference in collaborative reasoning within their groups. • The overall conclusion is that there is a gap between students' metacognitive knowledge of collaborative scientific reasoning and their use of collaborative scientific reasoning skills and attainment of conceptual understanding.
Weight of evidence A (trustworthiness in relation to study questions)	Medium Not any higher because of remarks at M4, and general lack of information on the rigour of the qualitative component of the study.
Weight of evidence B (appropriateness of research design and analysis)	Medium-high Hypotheses I and IV are relevant for this review. Good sample size with careful selection method, from two schools only; design includes a control group, and collects benchmark data but slightly different from the intended outcome data; sizeable reliability for and validity for the data-collection method (apart from the MKA tool) and the high for data analysis.
Weight of evidence C (relevance of focus of study to review)	Medium-high Little information is provided about the actual group discussions; the group discussions are a major but not distinct component of the intervention; the measures focus directly on students' understanding of evidence over some breadth; the situation of the study is classroom based but in a largely white and middle-class setting.
Weight of evidence D (overall weight of evidence)	Medium-high

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 undergo the intervention and generate collaboratively a report for two laboratory activities. The domain-specific knowledge for one activity is low, for the other high. • Reasoning strategies in interactions between pairs are video-recorded, and in individual and joint written products are collected. The types of reasoning strategies resulting in conceptual change are identified. For paper 2, no interviews are used, and three pairs are involved. The types of reasoning strategies used are 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; c. coordinating all mutually consistent knowledge propositions into a coherent model. RQ2: A comparison between the reasoning strategies employed in activities with low and high domain-specific demands respectively, is not really made. 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); d. for support (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 with 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 (appropriateness of research design and analysis)	Low Only RQ4 is relevant for this review. This RQ is part of the non-evaluative component of this interpretative study. For this review generalisations of findings are required, which this study with its small sample, no comparison and emphasis on contextual validity does not intend.
Weight of evidence C (relevance of focus of study to review)	Medium-low The nature of the sample, although carefully constructed and justified, is not representative. The variable of small-group discussion is an integral part of the intervention (collaborative report-writing). The measures (reasoning strategies) indicate understanding of evidence directly, but the progress in this type of understanding is not part of the pre-post design. The focus is on several aspects of understanding of evidence, and the classroom setting is naturalistic.
Weight of evidence D (overall weight of evidence)	Medium-low

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	Pupils from 2 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 were 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	Verbal data was not analysed but used as exemplars to support computer data. Statistical for computer data.

data, including details of checks on reliability and validity	• 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 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	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 high 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 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 type of search strategies used by successful students were different than 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, dynamically within the environment their diagnoses increased. Confidence is a true indicator of students' diagnostic accuracy. RQ 3 • There were some differences in tutoring strategies between a teacher and a GS.
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 (appropriateness of research design and	Low Small sample size and no sampling method. No control or pre-post testing. No information concerning reliability or validity of data-collection. Some validity check for data analysis.

analysis)	
Weight of evidence C (relevance of focus of study to review)	Medium Not very representative sample as based in an all girls, private school. The focus was on the computer cues, rather than on the discussion. The measures of understanding and their breadth were good. The classroom situation was representative of typical T & L.
Weight of evidence D (overall weight of evidence)	Medium-low

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: Stated 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 positive 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 were reported. • Unpaired t-tests were used to compare pre-test scores for intervention and control groups and determine equivalence. • Unpaired t-tests were used to compare post-test scores for intervention and control groups and 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 Chi-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 previously 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 (appropriateness of research design and analysis)	Medium-low The study uses a substantial sample without justifying the use of the school(s?). A non-equivalent control group is used and the pre-post testing is done for logical thinking but not specifically understanding of evidence. The validity and reliability of the data-collection tool was secured but no detail is provided for the reliability and validity of the data analysis method.
Weight of evidence C (relevance of focus of study to review)	Low The study hardly focuses on the effect of small-group discussion and this aspect cannot be isolated for evaluation from the rest of the intervention. The measures do not include tape-recorded student discussions and focus on the improvement of logical reasoning skills rather than understanding of evidence. The classroom situation is representative, although little detail exists for the nature of the students involved.
Weight of evidence D	Medium-low

(overall weight of evidence)	

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. Pupils 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 were collected (see later sections). This study focuses on analysis of discourse. Actual sample: Nine classes with an average of 26 pupils implies a sample size of around 230 pupils.
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: pupil 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	• Pupils 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. • Pupils initially found whole class discussion and debate about reaching a consensus confusing, but did ultimately arrive at an agreed scientific view. • Pupils enjoyed planning the investigation.

	• Pupils used explanations to scaffold their discussions, particularly to provide reasons for their proposals. Pupils also discussed explanations. • Pupils stayed focused on discussion tasks. • Pupils were able to use their previous everyday experience to inform planning of investigations. Pupils 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 a significantly greater number of pupils 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 The RQ relevant for this review is: What is the role of explanations in scaffolding group discussions? The sample size large but the sampling method could be more specific; the cross sectional study design misses a control group or benchmark data on understanding of evidence; few checks for reliability and validity in data-collection and analysis are reported.
Weight of evidence C (relevance of focus of study to review)	Medium-low The nature of the intervention is highly representative of small-group discussions; the independent variable and the dependent variable in the relevant aspect of the study are reversed compared with the review; the situation is representative of classroom learning.
Weight of evidence D (overall weight of evidence)	Medium-low

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 US 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 in which dyads of learners, grouped according to ability (high/high: low/low: high/low) worked 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 The sample size was large and method of sampling was carefully balanced. A control group was employed, although no pre-testing was carried out. The issue of reliability and validity of data-collection and analysis were sufficiently addressed.
Weight of evidence C (relevance of focus of	Medium-low The type of SGD expected from the students was representative but the main focus of the study was on the design of the computer software

study to review)	and the measures were therefore less relevant to this review. The breadth of the measures were sufficient and the classroom situation was representative of typical T & L situations.
Weight of evidence D (overall weight of evidence)	Medium

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. Interaction 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 Only the last cycle of this formative evaluation aimed at modifying the Belvedere tool is relevant for this review. The sample size is small but unspecified without a sampling method; the design does not involve a control group, and has no pre-post component; extremely few details are provided for the reliability and validity checks for data-collection and analysis.
Weight of evidence C (relevance of focus of study to review)	Medium No details are provided to judge the relevance of the nature of the group discussion; the intervention uses group discussions as a major, but not separable, component; where the measures (computer work) combine gauging procedural skills and understanding of evidence, of which a wide range of aspects has been measured; the situation (pairs of students on adjacent computers) is not representative.
Weight of evidence D (overall weight of evidence)	Medium-low

Tao P-K (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	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, though 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 work 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 to 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 mean score of 17.75 and 18.26 and p = 0.87. Validity of feedback instrument with varying alternative solutions 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	Medium Indicators for problem-solving skills not clearly stated. Reported abilities (e.g. meta-cognition) unrelated. Reliability and validity of data-

to study questions)	collection methods and analysis methods 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, no comparison control group, limited and largely descriptive information on problem-solving skills does not allow for conclusions on effects of the intervention.
Weight of evidence C (relevance of focus of study to review)	Medium-low The nature of the sample is slightly atypical (highly motivated students); the small-group discussion technique is not a major part of the intervention centred around varied feedback; the measure indicates problem-solving, not understanding of evidence, let alone a breadth of this understanding; the testing situation in class is representative.
Weight of evidence D (overall weight of evidence)	Medium-low

1. Tolmie A, Howe C (1993) Gender and dialogue in secondary school physics. *Gender and Education* **5:** 191-209. 2. Howe C, Tolmie A, Anderson A (1991) Information technology and group work in physics. *Journal of Computer Assisted Learning* 7: 133-143.	
Country of study	UK
Details of researchers	Researchers at the University of Strathclyde, Glasgow, funded by the Economic and Social Research Council
Name of programme	Not applicable
Age of learners	12 to 15
Type of study	Evaluation: researcher-manipulated
Aims of study	To investigate whether established gender differences in expression of opinion have a substantial impact on the exchange of opinions between pupils engaged on a science task. The consequences for understanding of exchanging ideas while making joint decisions and whether gender composition of groups made a difference to learning and how decisions were reached.
Summary of study design, including details of sample	Identical pre- and post-intervention test with four 'explanation' tasks were carried out. Small-groups were composed of three differently gendered types of pairs. Interactions of pairs during the three intervention phases were observed. Compared pre-post test scores for differently gendered pairs and interaction patterns for differently gendered pairs. Actual sample: 82 at start, data were used from 73 pupils available to do the post-test.
Methods used to collect data	• Curriculum-based assessment • Observation: 12-13 indices of on-task activities by videotaping dialogues • Psychological test • Computer record of joint predictions
Data-collection instruments, including details of checks on	• Verbatim tasks (as computer screens) for comparing original responses, constructing a joint prediction, input this prediction and comparison with correct solution are all provided. Checks on reliability: Made on pre-intervention responses and provided 90% inter-judge agreement.. Test for scoring of dialogues gave 81%

reliability and validity	inter-judge agreement. Multiple tasks aimed at the same underlying concepts increase reliability. Checks on validity: Scoring of predictions and explanation problem responses had been used previously by authors and disseminated (Anderson *et al* 1990). Triangulation (video records and computer logs) increase the validity of the collection method.
Methods used to analyse data, including details of checks on reliability and validity	• Mean scores for each pupil on the first test deducted from mean score on the second yielded a measure of explanation change. • Patterns of group interaction were analysed by 'causal analysis' (Blalock, 1972). • Comparison of pre-post test scores for participants in male, female and mixed groups • Correlations between change in test scores and (i) membership of gendered groups, (ii) the amount of initial dissimilarly within groups and (iii) the amount of discussion of explanatory factors within groups. • Calculation of mean scores for pre and post test (values for means provided but no sd) • Significance testing and analysis of variances for differences in these scores • Causal analysis' (interesting) based on correlations between all possible pairs, and statistically different relationships, of interaction characteristics and their sequence in time Checks of reliability: For causal analysis, use published method of Blalock (1972). Checks on validity: None
Summary of results	• The intervention caused an overall significant improvement of individual explanatory understanding: means from 1.13 to 1.47 ($F=5.49$, $df=1.71$, $P<0.05$). • This change does not differ for members of female, male or mixed groups ($F = 2.14$, $df = 2.70$, p ns). • The change correlates positively with the initial dissimilarity of the group members ($r = +0.19$, $p = 0.05$). • Interactional styles differ for male, female and mixed pair interactions, although they yield the same improvement of understanding. • Male pairs learn most when attending to differences in predictions and feedback lead to discussion of factors at work, and taking these into account be re-constructing their explanations. • Female pairs learn by identifying but ignoring differences in predictions and feedback. Though no on-task adjustment of ideas, searching for (common) explanations across tasks improved understanding. • Mixed pairs also avoided identified conflicting explanations, mainly by taking turns in documenting understanding. No explicit coordination of ideas and evidence (as in all-male), and no co-ordination between ideas relevant to different problems (as in all-female).
Conclusions	• That both interaction style and manner of progress through a task do differ as a function of a group's gender composition. The actual nature of the observed patterns of interaction suggests that the major source of difference is the social effect of conceptual conflict; the process of opinion exchange was central. • Overall, the results suggest that group-orientated software which encourages joint decisions would be worth developing in the teaching of physics. • The software could be improved, not so much to cater for the male pairs since the software worked well for them as it stood. Rather, to adapt to the apparent requirements of the female and mixed pairs which were weak at predictions. Suggestions are made by the authors for ways that could assist predictive discussion: for example, by presenting on screen a range of possible predictions and requiring one to be selected.
Weight of evidence A (trustworthiness in relation to study questions)	Medium-high The reliability and the validity for data scoring have been checked thoroughly, slightly less so for data analysis. The experimental setting prevents generalisation.
Weight of evidence B (appropriateness of	Medium-high The sample size was sufficient and care was taken with balancing group composition for ability and gender. Thorough pre- and post-testing

research design and analysis)	was carried out and benchmark data obtained. The reliability for data-collection and data analysis was satisfactory.
Weight of evidence C (relevance of focus of study to review)	Medium Somewhat representative sample because only from one school. The SGD is the prime focus of the intervention. The measures were highly appropriate for testing understanding of evidence, although there was a somewhat narrow range of evidence (prediction and explanation considered). The out-of-class setting was not representative of class learning.
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 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 other 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	No details are given and there was no formal analysis. Quotes from students are included to support conclusions.

data, including details of checks on reliability and validity	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, only post-intervention data collected, no checks on reliability and validity of data-collection instruments or analysis.
Weight of evidence C (relevance of focus of study to review)	Low Small-group discussions wrapped up in intervention, little information on students' understanding of evidence but was carried out in a representative classroom situation
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, audio-taped 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 • Audio-tapes 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 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-ratter 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-ratter reliability scores calculated. Checks on validity: No details were 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 for generating interest and connecting to out-of-school life experiences. • Student learning was aided by having students work in pairs and or 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 suggests that many may relate to specific properties the context of the intervention. They also note that many of the teachers and students were very enthusiastic about the programme, again suggesting caution 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 RQs 4 and 5 are relevant for this review. The design uses a sizeable sample without a clear sampling methods, a distinct control group and pre-post collection of data on argumentation. The reliability and validity of the data-collection method was hardly mentioned and some detail was provided for the reliability for data analysis only.
Weight of evidence C (relevance of focus of study to review)	Medium The small-group discussions reported are a focus subsidiary to the written work and are atypical as they are short. The study focuses on evaluating the Genetic Revolution unit of which small-group discussion is an integral component. The measures for argumentation largely focused on written work but described a range of understanding of evidence. The situation of the study was part of the classroom teaching.
Weight of evidence D (overall weight of evidence)	Medium