REVIEW

April 2005

A systematic review of the nature of small-group discussions aimed at improving students' understanding of evidence in 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-Centre 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
Alison Robinson	Department of Educational Studies, University of York, UK

REVIEW GROUP MEMBERSHIP

Dr Judith Bennett	Department of Educational Studies, University of York, UK (Review Group Co-ordinator)
Martin Braund	Department of Educational Studies, University of York, UK
Dr Bob Campbell	Department of Educational Studies, University of York, UK
Nick Daws	Inspector at the Office for Standards in Education (Ofsted) and Science Inspector for Staffordshire LEA, UK
Steve Dickens	Head of Physics, Dixons City Technology College, Bradford, UK
Alison Fletcher	Head of Science, Huntington School, York, UK
Nichola Harper	Head of Chemistry, Aldridge School, Walsall, UK
Dr Sylvia Hogarth	Department of Educational Studies, University of York, UK (Review Group Research Fellow)
Professor John Holman	Department of Chemistry, University of York, UK, and Director of the Science Curriculum Centre, University of York
Declan Kennedy	University College, Cork, Ireland, and science textbook author
Dr Ralph Levinson	Institute of Education, University of London, UK
Fred Lubben	Department of Educational Studies, University of York, UK (Review Group Research Fellow)
Alyson Middlemass	Assistant Head Teacher and Head of Science, Elizabethan High School, Retford, UK
Professor Robin Millar	Department of Educational Studies, University of York, UK
Christine Otter	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-Centre Review Group for English

ACKNOWLEDGEMENTS

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

The EPPI-Centre 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 time funded by the Higher Education Funding Council for England (HEFCE) for two of the core team members, Judith Bennett and Bob Campbell.

CONFLICTS OF INTEREST

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 in its pilot phase (at time of writing) which will be advocating the use of small-group discussions. A number of members of the RG (Judith Bennett, Bob Campbell, John Holman, Robin Millar) were involved in the development of the *Salters* courses (*Science: the Salters Approach, Science Focus, Salters Advanced Chemistry, Salters Horners Advanced Physics*), all of which advocated the use of small-group discussions as one of a range of student-centred approaches in teaching.

LIST OF ABBREVIATIONS

ANOVA	Analysis of variance
ASE	Association for Science Education
BEI	British Education Index
CBI	Computer-based instruction
CLE	Computer-supported learning environment
DfEE	Department for Education and Employment
DfES	Department for Education and Skills
EPPI-Centre	Evidence for Policy and Practice Information and Co-ordinating Centre
ERIC	Educational Resources Information Centre
HEFC	Higher Education Funding Council for England
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	Postgraduate Certificate in Education
POLS	Perspectives on learning science

PSE	Personal and social education
QCA	Qualifications and Curriculum Authority
RCT	Randomised controlled trial
RG	Review Group
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
WoE	Weight of evidence

This report should be cited as: Bennett J, Lubben F, Hogarth S, Campbell B and Robinson A (2005) A systematic review of the nature of small-group discussions aimed at improving students' understanding of evidence in science. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

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

TABLE OF CONTENTS

SUMMARY

Background

This review 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 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*
- 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 (DfEE, 1998) in England and Wales), where small-group discussions are seen to play an important role

Aims

The principal aim of the review is to explore the nature of small-group discussions aimed at improving students' understanding of evidence is science.

The review is the third of three reviews focusing on aspects of small-group discussion work in science lessons.

Review questions

The review question is as follows:

What is the nature of small-group discussions aimed at improving students' understanding of evidence in science?

The question has emerged from the initial question identified by the Review Group on small-group discussion work:

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

This resulted in an in-depth review (Bennett *et al.*, 2004) which looked at the question:

What is the evidence from evaluative studies of the effect of small-group discussions on students' understanding of evidence in science?

Using an updated version of the systematic map developed for the first review, a second in-depth review (Hogarth *et al.*, 2004) addressed the question:

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

This third in-depth review also explores an area of the updated systematic map.

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 studies which explored the nature of small-group discussions aimed at improving students' understanding of evidence, and focused on the nature of small-group discussion work as a key discrete variable in their data analysis.

Methods

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

- *Searching and screening*: developing criteria by which studies are to be included or excluded in the review, searching (through electronic databases and by hand) for studies which appear to meet these criteria, and then screening the studies to see if they meet the inclusion criteria

- *Keywording and generating the systematic map*: coding each of the included studies against a pre-agreed list of characteristics which is then used to generate a systematic map of the area, whereby 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

In addition, and very importantly, this review has attempted to draw on the recently published guidance and framework for assessing research evidence in qualitative research studies (Spencer *et al.*, 2003). Drawing on this guidance was seen as crucial for this review, as all the papers included in the in-depth review reported on qualitative studies. In order to draw on the guidance, a range of additional questions was developed and integrated into the data extraction questions in the EPPI-Centre tool. Full details are given in the main report, and a copy of the additional questions is included as Appendix 2.5.

Systematic map

The number of studies identified through the searching and screening established that small-group discussions were being used in a variety of ways in science lessons. However, in many of the studies 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.

There were 94 studies 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 USA, 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; less relates to students’ attitudes.
- A diversity of measures was used to assess effects on understanding and attitude.
- Very little research has been done on small-group discussions in relation to the teaching of chemistry.
- Typical small-group discussions 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.
- There were 28 studies with experimental designs, of which 12 were randomised controlled trials (RCTs).

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

In-depth review

Nineteen studies were included in the in-depth review, which focused on the nature of small-group discussion work aimed at improving students' understanding of evidence.

The consolidated evidence from the review draws primarily on the findings from studies weighted as medium-high and, to a lesser extent, as medium in terms of their overall quality.

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

Findings on nature of small-group discussions

Although there is considerable variety in the detailed research questions and discussion topics used to promote small-group discussion, there is a high degree of consistency in the findings and conclusions. In general, students often struggle to formulate and express coherent arguments during small-group discussions, and demonstrate a relatively low level of engagement with tasks. The review presents *very strong evidence of the need for teachers and students to be given explicit teaching in the skills associated with the development of arguments and the characteristics associated with effective group discussions*. Five of the seven highest quality studies in the review make this recommendation. There is also *good evidence* to confirm the findings of other reviews on small-group discussions (Bennett *et al.*, 2004; Hogarth *et al.*, 2004) on the desirability of the stimulus used to promote discussion involving both internal and external conflict, i.e. where a diversity of views and/or understanding are represented within a group (internal conflict) and where an external stimulus presents a group with conflicting views (external conflict).

There is *good evidence* on group structure. Not all studies addressed this aspect, but, where advice is offered, it tends to indicate that groups should be specifically constituted such that differing views are represented. There is also evidence to suggest that assigning managerial roles to students (e.g. reflector, regulator, questioner, explainer) as suggested in collaborative learning theory is likely to be counter-productive for poorly-structured tasks. Some evidence is also presented which suggests single-sex groups may function better than mixed-sex groups, although overall development of understanding is not affected by group composition. Group leaders also emerge as having a crucial role: those that were able to adopt an inclusive style, and one which promoted reflection, were the most successful in achieving substantial engagement with the task. An alienating leadership style generates a lot of off-task talk and low levels of engagement.

The review presents some evidence that small-group discussion work does improve students' understanding and use of evidence. Whilst this was not the main focus of the review, all the included studies present some evidence in this area, as improvement in use of evidence was one of the reasons for using small-group discussions. The effects of small-group discussions on students' understanding of evidence has been explored in more detail in other reviews (Bennett *et al.*, 2004; Hogarth *et al.*, 2004).

Findings on research strategies adopted to explore aspects of small-group discussion work

A number of similarities emerged in the approaches adopted in the studies. They tended to make use of opportunistic samples, drawing on the researchers' personal contacts. Experimental designs are not often used, although studies often made comparisons between discussion groups in the same class or within a discussion group. Data collection methods typically involve audio- and/or video-recordings, with analysis and reporting drawing heavily on extracts from recorded dialogue. Whilst approaches to gathering data are seldom justified in any detail by the authors, sound procedures appear to be introduced to check the reliability of the data analysis and to present the findings in a way which makes them trustworthy.

A key difference which has emerged concerns the two contrasting approaches to data analysis, with some studies developing grounded theory from the data, and others drawing on existing models to structure their analysis.

Strengths of the review

The review has a number of strengths:

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

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

- The review has deliberately focused on synthesising the evidence from the studies rated as medium-high in quality. (No studies were rated as high in quality.)

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

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

Limitations of the review

The review has one principal limitation. Although the studies in the in-depth review share a number of similar characteristics at the broad level, there are substantial differences at the detailed level. For example, there is considerable variety in the specific research questions, the topics used for the discussion tasks, and in the use and interpretation of key terms relating to *evidence* and *understanding of evidence*. However, the effect of this limitation was minimised by focusing the in-depth review on studies of medium-high quality. (No studies were rated as high quality.)

Implications for policy

Current policy strongly advocates the use of small-group discussion work. Whilst the main focus of this review was to establish *how* small-group discussions were being used in science lessons, it also yielded evidence of some potential benefits in terms of helping students develop their skills in formulating arguments; hence the review does indicate that there could be benefits in pursuing such a policy. However, it is clear from the review that small-group discussion work needs to be supported by the provision of guidance to teachers and students on the development of the skills necessary to make such work effective. Thus, some form of professional development training for teachers would appear to be highly desirable to provide them with guidance on how to maximise the effectiveness of small-group discussions.

Implications for practice

The review suggests that small-group discussion work can provide an appropriate vehicle for assisting students in the development of ideas about using evidence and constructing well-supported arguments. Thus teachers should be encouraged to incorporate such discussions into their teaching, provided that appropriate support is offered to help them develop the necessary skills (see section 5.3.1). Gathered additional research data on their use and effects would also be very important.

Implications for secondary research

The review indicates that the most useful form of secondary research which could be pursued would be to look at methods used to analyse student discourse to establish similarities and differences in existing frameworks and frameworks emerging from grounded theory.

Implications for primary research

One particularly strong feature which has emerged from the work undertaken for this review and the others on small-group discussion (Bennett *et al.*, 2004; Hogarth *et al.*, 2004) 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 to implement good practice. 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, linked to a coherent analysis framework drawing on the findings of the secondary research proposed above.

Other aspects

This review made use of an enhanced data-extraction tool developed by the Review Team to address the fact that the reports of the studies presented a significant amount of qualitative data.

The enhanced data-extraction tool asks for specific details of relevance to qualitative studies to be entered into the database EPIC, using EPPI-Reviewer, the EPPI-Centre software. These details relate to the design of the study, important features of the data collection, important features of the analysis, and ethical considerations.

Overall, the tool was found to be very helpful in systematically identifying and recording details of studies which might not have been captured in the standard data-extraction tool. In addition, the enhanced data-extraction tool served to identify areas where qualitative studies provided good and appropriate detail, and areas where more detail would have been helpful. A particularly positive feature to emerge concerned the steps taken to increase the reliability and trustworthiness of data analysis in qualitative studies.

1. BACKGROUND

1.1 Aims and rationale for current review

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

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

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

This review led to a systematic map of the area and a first in-depth review of studies addressing the following question:

What is the evidence from evaluative studies of the effect of small-group discussions on students' understanding of evidence in science?

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

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

It became apparent in the first and second reviews that small-group discussions were used in a variety of ways in science lessons, and that a consolidation of descriptive evidence of the characteristics of small-group discussions would be desirable. This review therefore addresses the following question:

What is the nature of small-group discussions aimed at improving students' understanding of evidence in science?

Part of the process of undertaking the review involved drawing on the recently-published guidance and framework for assessing research evidence in qualitative research studies (Spencer *et al.*, 2003). In order to draw on the guidance, a range of additional questions was developed and integrated into the data-extraction questions.

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 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, happening in real time and, most usually, face to face) or *asynchronous* (that is, not happening in real time and 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- to 16-year-olds, has resulted in the development of a new science course for this age range, *21st Century Science*, which is discussed on its website (http://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- to 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*: 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

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

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. 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 is made to small-group discussion work, although, by implication, it must have been taking place. It was therefore decided that, for the purposes of the research review question, this area of work would be excluded unless reference was made to the use of specific discussion tasks and their effects.

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

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

Envoying: This technique also involves students working in two groups. In the first group, they discuss a common task, which differs for each group. Groups then 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, academics, school governors and parents.

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(s)
- the development of inclusion and exclusion criteria
- the development of review-specific keywords
- the identification of the focus for the in-depth review
- the content of the report(s)

The Review Group has met regularly to monitor and discuss progress, and to advise and guide the core team.

School students are also a key constituency group. While it is impractical to invite them to attend Review Group meetings, they have been involved in commenting on the findings of the review.

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

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 were 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. 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

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

The EPPI-Centre systematic review methods were followed for searching, screening and including (or excluding) studies in the map, and in applying the EPPI-Centre keywording sheet and keywording strategy (EPPI-Centre, 2002a, 2002b), supplemented by review-specific keywords, to studies. Extracting data and making quality assessments of studies included in the in-depth review was also carried out according to EPPI-Centre procedures and using EPPI-Centre software (EPPI-Centre, 2002c). In addition, questions developed from the recently-published guidance and framework for assessing research evidence in qualitative research studies (Spencer *et al.*, 2003) were integrated into the data-extraction questions.

The systematic map was based on studies identified in the second review (Hogarth *et al.*, 2004). Studies were included if they satisfied the following criteria:

- They were about the use of small-group discussions in science lessons.
- They involved groups of two to six students.
- They involved a substantive, structured discussion task of two minutes' duration or more.
- They illustrated how small-group discussions are being used.
- If they focused on learning outcomes, addressed aspects of students' understanding in science, or addressed aspects of students' attitudes to science.
- They were empirical studies of the following types: descriptive, exploration of relationships, evaluation (naturally-occurring and researcher-manipulated), reviews (systematic and non-systematic).
- They were about students in the 11–18 age range.
- They had been undertaken in the period 1980–2003.

- They were published in English.

Justification of these inclusion criteria may be found in the report of the first review on small-group discussion work (Bennett *et al.*, 2004).

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

2.2.2 Identification of potential studies: search strategy

The search strategy for this in-depth study was to use the relevant studies already identified in the first in-depth strategy and to update them to include papers published in 2003. The same methods of electronic and handsearching, and use of personal contacts were employed.

2.2.3 Screening studies: applying inclusion and exclusion criteria

The Review Team updated the database system (which uses EndNote software) for keeping track of, and coding, papers found during the review. Full details of the process may be found in the report of the second review (Hogarth *et al.*, 2004).

2.2.4 Characterising included studies

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

The second review (Hogarth *et al.*, 2004) produced a systematic map of the research in the area using the generic and review-specific keywording sheets. This is replicated in Chapter 4 in the form of narrative and mapping tables covering 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

The procedures required for quality-assurance purposes have been described in Hogarth *et al.* (2004). This includes validity measure for the inclusion criteria and keyword descriptors, and inter-rater reliability levels for screening and keywording.

2.3 In-depth review

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

The purpose of in-depth reviewing is to describe the characteristics of studies in more detail, and 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 follows:

What is the nature of small-group discussions aimed at improving students' understanding of evidence in science?

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

1. Exclusion on study focus: that is, the primary focus of the study was not on the nature of the small-group discussion – in other words, how they were *used* by the teacher.

2. Exclusion on study focus: that is, the study does not focus on the nature of the discussions.

3. Exclusion on place of publication: that is, the study is not published in a peer-refereed journal.

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', '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). In addition, questions developed from the recently-published guidance and framework for assessing research evidence in qualitative research studies (Spencer *et al.*, 2003) were integrated into the data-extraction questions.

The process used to develop this enhanced data-extraction tool was as follows:

- A mapping exercise was undertaken in which points/questions in EPPI-Reviewer and the qualitative research framework were mapped against each other. This was done independently by four members of the team (Bennett, Campbell, Hogarth, Lubben); results were then compared and moderated.
- Areas of overlap and areas of difference were noted.
- Questions were developed by three team members (Bennett, Campbell and Lubben) who addressed aspects of relevance in qualitative studies which were not covered by questions in EPPI-Reviewer.
- These questions were then integrated into appropriate sections in EPPI-Reviewer (Sections E, I, J and M), with a small number of questions asked on EPPI-Reviewer being identified as not suited to qualitative studies.
- A trial version of the enhanced data-extraction tool was piloted by four team members (Bennett, Campbell, Hogarth, Lubben) on three papers, and results were compared. Final modifications were then made to the tool.

Appendix 2.5 shows the additional questions developed, the sections where they were inserted into EPPI-Reviewer, and the questions in EPPI-Reviewer deemed inapplicable to qualitative studies.

The focus of the additional questions informed decisions made about the criteria for assigning weights of evidence (see the following section and Appendix 2.6).

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

Once data have been extracted from the studies, the next step in the review process is to assess the quality of the studies and the weight of evidence they present in relation to the in-depth review question. The EPPI-Centre data-extraction procedures identify three quality levels – 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, these were refined into five categories: high, medium-high, medium, medium-low and low.

The categories are as follows:

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

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

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

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

For category A, a judgement of quality within the EPPI-Centre data-extraction guidelines (EPPI-Centre, 2002d) was used (M.11).

Judgements of weighting in categories B and C are based on the quality of the reported research solely in relation to the in-depth review question. Appendix 2.6 shows how the Review Team interpreted the appropriateness of the research design and analysis (category B) through five aspects: the sampling frame; the actual sample, the context of the small-group discussion, the data collection methods, and the data analysis. 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 2.6 shows how the relevance of the research focus of the study (category C) has been weighted through five aspects: the focus of the intervention, the focus of the study, the measures used to assess the nature of the discussion, the breadth of discussion reported and the representativeness of the study situation (learners in the classroom). Again, each of these aspects has three level descriptors with weighting 3, 2 or 1 in decreasing appropriateness. The sum total of the weighted aspects determines the overall weight of category C in the same way as explained for category B above.

The total weighting for category D was constructed by the Review Team by combining 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 table (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 members. Only one characteristic considered important was not included in this tool: the ‘details of the researchers’ and this information is included in the summary tables. These tables were edited by one Team member for consistency of terminology, depth and detail. 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, commonalities of, and differences between, studies were identified for methodological aspects of the studies on the basis of these reports. The latter resulted in the judgement of ‘weight of evidence A’. For the synthesis of the appropriateness of the studies’ research design and analysis (weight of evidence B), the five characteristics listed in weight of evidence B were used as organisers. The same was the case for the synthesis of the relevance of the focus of the studies (weight of evidence C). This synthesis method necessitated a continuous consultation between two 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 *medium-high* and, to a lesser extent, as *medium*, as summarised above. (No studies were rated as *high*.) Findings from studies weighted as *medium-low* or *low* have not been considered due to weakness in design and analysis compromising the strength of the evidence presented.

2.3.5 In-depth review: quality-assurance process

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

3. IDENTIFYING AND DESCRIBING STUDIES: RESULTS

3.1 Studies included from searching and screening

Figure 3.1 provides a summary of the number of papers and studies involved at various stages of the filtering process. The process of searching yielded 2,246 papers, of which 249 were identified by updating the period covered to include 2003. An additional 44 papers were identified through handsearching or personal contacts; thus the review handled a total of 2,290 records. After de-duplication and the first round of screening 391 papers remained for possible inclusion. Hard copies of only 12 papers (3%) were unobtainable. After second screening, 119 papers remained for inclusion in the review. Papers reporting on the same study were identified as described at the beginning of section 2.2. The 119 papers were found to report on 94 studies, of which 19 were included in the in-depth review.

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

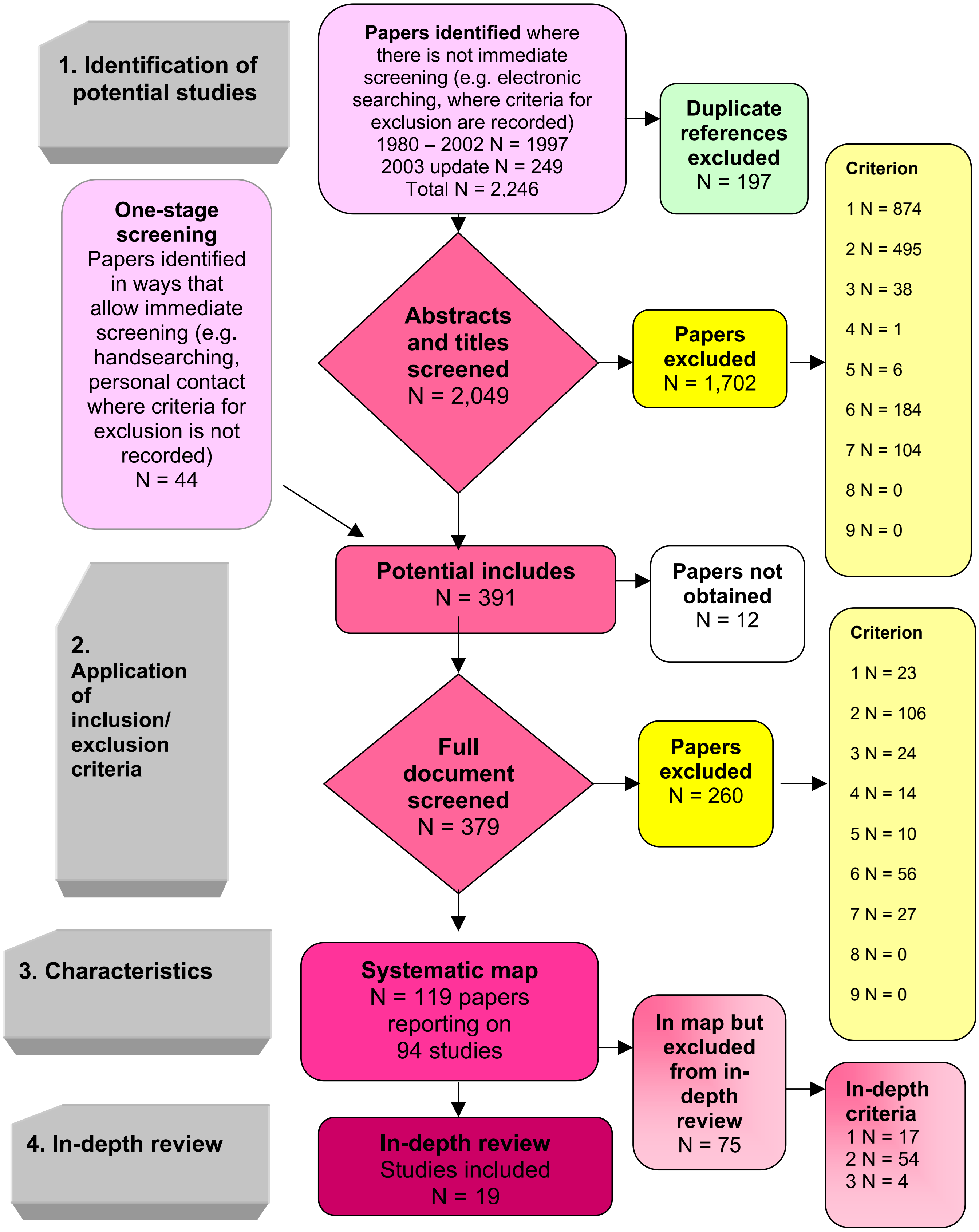


Table 3.1: Publication date of studies included in the systematic map (N = 94, mutually exclusive)

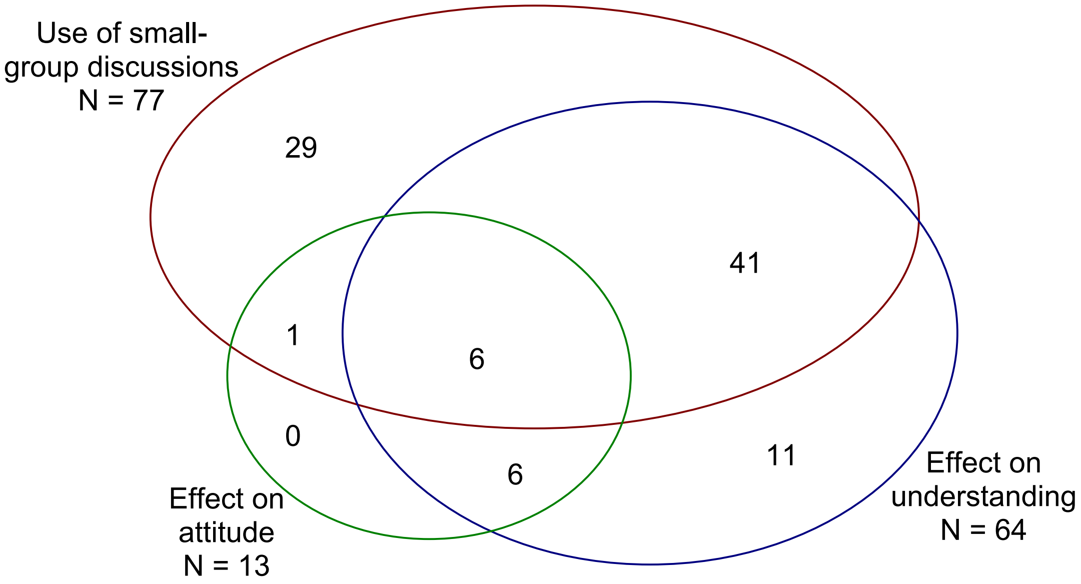
Publication period	Number of studies	%
1980–1985	1	1
1986–1991	5	6
1992–1997	38	40
1998–2003	50	53
Total	***94***	***100***

Table 3.1 indicates that the research activity in the review area has been minimal up to ten years ago and has 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

The review question has three components. The first component focuses on the process of what takes place during small-group discussions, in short the *nature 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: Characteristics of the included studies



NB: Venn diagram is not to scale.

Figure 3.2 indicates the focus of the 94 studies included in the review. Not surprisingly, the majority of studies (77) report on the process of small-group discussions, although only 29 of these solely report on this aspect. Just over half these studies (41) also report on the effect on students' understanding of science. A total of 64 studies report on the effect on students' understanding of science, with 11 of these dealing only with this aspect. A small number of studies (13) report on the effect of small-group discussions on students' attitude to science, with about half of these (6) reporting on all three aspects of the review question.

3.3 Identifying and describing studies: quality-assurance results

The results of the quality-assurance processes for searching, screening and keywording are as follows:

The inter-screener reliability as measured by the frequency method and the Cohen's Kappa method is shown in Table 3.2. The Cohen's Kappa method has the advantage of compensating for chance agreement.

Table 3.2: Inter-screener agreement (include-exclude) for first and second screening

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

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

As a result of the screening process, five new studies were added to those included in the earlier review of small-group discussions (Bennett *et al.*, 2004). These studies were keyworded independently by two team members, with an inter-coder agreement of 92.2%.

4. IN-DEPTH REVIEW: RESULTS

4.1 Selecting the studies for the in-depth review

This chapter reports on the in-depth review. It looks in detail at the results of a subset of the studies in the systematic map which was chosen because these studies are about the nature of small-group discussions aimed at improving students' understanding of evidence in science.

The application of the exclusion criteria specified in section 2.3.1 resulted in 19 studies for the in-depth review. A substantial number of studies met the criteria for in-depth review.

Studies included in the in-depth review

1. 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.
2. 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.
3. Hogan K (1999a) Sociocognitive roles in science group discourse. *International Journal of Science Education* **21:** 855–882.
4. 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.
5. Jiménez-Aleixandre MP, Pereiro-Muñoz C (2002) Knowledge producers or knowledge consumers? Argumentation and decision making about environmental management. *International Journal of Science Education* **24:** 1171–1190.
6. Jiménez-Aleixandre MP, Rodriguez AB, Duschl RA (2000a) 'Doing the lesson' or 'doing science': argument in high school genetics. *Science and Education* **84:** 757–92.
7. 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.
8. 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.
9. Kurth LA, Anderson CW, Palincsar AS (2002) The case of Carla: dilemmas of helping all students to understand science. *Science Education* **86:** 287–313.
10. Lajoie SP, Lavigne NC, Guerrera C, Munsie SD (2001) Constructing knowledge in the context of BioWorld. *Instructional Science* **29:** 155–186.
11. Meyer K, Woodruff E (1997) Consensually driven explanation in science teaching. *Science Education* **81:** 173–192.
12. Palincsar AS, Anderson C, David YM (1993) Pursuing scientific literacy in the middle grades through collaborative problem solving. *Elementary School Journal* **93:** 643–658.
13. 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.

14. 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.
15. Tao PK (2001) Developing understanding through confronting varying views: the case of solving qualitative physics problems. *International Journal of Science Education* **23:** 1201–1218.
16. Tolmie A, Howe C (1993) Gender and dialogue in secondary school physics. *Gender and Education* **5:** 191–209.
17. 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.
18. 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.
19. 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 more than one linked papers. One paper was selected as the lead paper for each study, but data in both or all three papers were drawn on for data-extraction purposes. The linked papers are as follows:

- Tolmie and Howe (1993) and *Howe, Tolmie and Anderson (1991)
- Keys (1997) and *Keys (1995)
- Roth and Roychoudhury (1992) and *Roth, Taylor and Roychoudhury (1994) and *Roth (1994)
- Tao (2001) and *Tao (2000b)

Additionally, there were links between Meyer and Woodruff (1997) and Woodruff and Meyer (1997) in that the latter report drew in very limited ways on the methodology and study reported in the former.

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

Summary of weights of evidence (WoE) judgements

The weights of evidence assigned to each of the 19 studies in the four categories are given in Table 4.1. Appendix 4.1 contains tables summarising for each study the key information used to inform judgements about the weights of evidence.

Additionally, specific indicators relating to the review question were developed to assist the process of making weight-of-evidence judgements in categories B and C, and an algorithm developed for combining weight-of-evidence judgements in categories A, B and C to arrive at the overall weight-of-evidence judgement in category D. These have been described in section 2.3.3, and an overview in the form of a table may be found in Appendix 2.6.

Applying these criteria and the algorithm allowed studies to be judged as high (H), medium-high (MH), medium (M), medium-low (ML) and low (L).

Table 4.1: Weights of evidence (WoE) assigned to each of the 19 studies in the in-depth review

	Study	**WoE A**	**WoE B**	**WoE C**	**WoE D**
1	De Vries *et al.*, 2002	M	M	MH	M
2	Finkel, 1996	MH	M	M	M
3	Hogan, 1999a	M	MH	MH	MH
4	Hogan, 1999b	M	M	M	M
5	Jiménez-Aleixandre and Pereiro-Muñoz, 2002	ML	ML	MH	M
6	Jiménez-Aleixandre *et al.*, 2000a	MH	MH	H	MH
7	Johnson and Stewart, 2002	ML	M	ML	ML
8	Keys, 1997	MH	MH	MH	MH
9	Kurth *et al.*, 2002	M	M	M	M
10	Lajoie *et al.*, 2001	ML	M	M	M
11	Meyer and Woodruff, 1997	M	ML	ML	ML
12	Palinscar *et al.*, 1993	M	ML	ML	ML
13	Richmond and Striley, 1996	MH	M	H	MH
14	Roth and Roychoudhury, 1992	M	MH	H	MH
15	Tao, 2001	ML	ML	ML	ML
16	Tolmie and Howe, 1993	MH	MH	M	MH
17	Tsai, 1999	M	ML	ML	ML
18	Woodruff and Meyer, 1997	ML	ML	MH	M
19	Zohar and Nemet, 2002	M	M	H	MH

Of the 19 studies, seven were rated medium-high (MH), seven were rated medium (M), and five were rated medium low (ML) overall, as indicated in table 4.2.

Table 4.2: Summary of overall weight-of-evidence judgements on studies

Medium-high (MH)	**Medium (M)**	**Medium-low (ML)**
Hogan, 1999a Jiménez-Aleixandre *et al.*, 2000a Keys, 1997 Richmond and Striley, 1996 Roth and Roychoudhury, 1992 Tolmie and Howe, 1993 Zohar and Nemet, 2002 (7)	De Vries *et al.*, 2002 Finkel, 1996 Hogan, 1999b Jiménez-Aleixandre and Pereiro-Muñoz, 2002 Kurth *et al.*, 2002 Lajoie *et al.*, 2001 Woodruff and Meyer, 1997 (7)	Johnson and Stewart, 2002 Meyer and Woodruff, 1997 Palinscar *et al.*, 1993 Tao, 2001 Tsai, 1999 (5)

It was pleasing to see that just under half the studies (seven) were judged to be medium-high overall in terms of the evidence yielded. This synthesis focuses on these seven studies, as they provide the strongest evidence base on which to make recommendations.

A note on the researchers

Of the 19 studies, six appeared to be undertaken by single researchers, three of whom appeared to be reporting on doctoral studies (Finkel, 1996; Hogan, 1999a and 1999b; Keys, 1997), although this was not stated explicitly in one of the reports (Hogan, 1999a). Additionally, the study by Johnson and Stewart (2002) was based on Johnson's doctoral work.

One study was completed by a post-doctoral researcher (Tao, 2001). Seven of the studies were undertaken by pairs of researchers (Jiménez-Aleixandre and Pereiro-Muñoz, 2002; Johnson and Stewart, 2002; Meyer and Woodruff, 1997; Richmond and Striley, 1996; Roth and Roychoudhury, 1992; Tolmie and Howe, 1993; Zohar and Nemet, 2002;). The remaining five studies were undertaken by teams of three or more researchers (De Vries *et al.*, 2002; Jiménez-Aleixandre *et al.*, 2000a; Kurth *et al.*, 2002; Lajoie *et al.*, 2001; Palincsar *et al.*, 1993).

Where details are given, or could be inferred from author details provided at the start of the reports, the majority of the authors appear to have been based in universities at the time of writing the reports. Johnson (Johnson and Stewart, 2002) was cited as being based both in a school and a university. 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 Palinscar *et al.* (1993), Roth in Roth and Roychoudhury (1992), and Nemet in Zohar and Nemet (2002).

On the basis of information provided, 12 studies report on work which had been externally funded: De Vries *et al.* (2002), Jiménez-Aleixandre *et al.* (2000a), Jiménez-Aleixandre and Pereiro-Muñoz (2002), Johnson and Stewart (2002), Kurth *et al.* (2002), Lajoie *et al.* (2001), Meyer and Woodruff (1997), Palinscar *et al.* (1993), Richmond and Striley (1996), Tao (2001), Tolmie and Howe (1993), and Woodruff and Meyer (1997). This comparatively high proportion of funded studies in the review is, perhaps, an indication of interest in the area.

It is worth noting that no details at all were provided of the status of the researchers (e.g. doctoral student, teacher-researcher, university-based academic) and their role in the study in six of the 19 papers. The absence of such information is a serious omission from a paper, as it is an important aspect of the context in which the study was undertaken.

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

This section compares certain characteristics of the studies selected for the in-depth review (country of study, science subject focus, age of learner) with those in the systematic map to establish the extent to which the studies in the in-depth review reflect those in the systematic map as a whole.

Countries of studies

Table 4.3 shows the countries in which studies selected for in-depth review were carried out. The majority of the studies were undertaken in North America (US and Canada), with others as detailed below. The proportion of studies undertaken in North America (63%) is rather larger than the proportion of studies in the map (40%). Those studies judged to be of medium-high overall quality (i.e. those which form the bulk of the evidence presented in the in-depth review) have been asterisked (*).

Table 4.3: Countries in which the studies selected for in-depth review were carried out (N = 19, mutually exclusive)

Country	Number of studies	Study
USA	8	Finkel, 1996 *Hogan, 1999a Hogan, 1999b Johnson and Stewart, 2002 *Keys, 1997 Kurth *et al.*, 2002 Palinscar *et al.*, 1993 *Richmond and Striley, 1996
Canada	4	Lajoie *et al.*, 2001 Meyer and Woodruff, 1997 *Roth and Roychoudhury, 1992 Woodruff and Meyer, 1997
Spain	2	*Jiménez-Aleixandre *et al.*, 2000a Jiménez-Aleixandre and Pereiro-Muñoz, 2002
France	1	De Vries *et al.*, 2002
Hong Kong/China	1	Tao, 2001
Israel	1	*Zohar and Nemet, 2002
Taiwan	1	Tsai, 1999
UK	1	*Tolmie and Howe, 1993

Subject focus

Nine of the 19 studies in the in-depth review focused on small-group discussions in Integrated Science lessons, six in Biology (including one where there was overlap between Biology and Earth Science), four in Physics and none in Chemistry.

This constitutes a higher proportion of Integrated Science lessons and Biology lessons, a similar proportion of Physics lessons and a lower proportion of Chemistry lessons in comparison with the studies in the systematic map. This difference may be due to the fact that understanding evidence comes to the fore in particular when discussing contentious issues, which often are related to biology: for example, genetic engineering or Human Immunodeficiency Virus (HIV)-Acquired Immunodeficiency Syndrome (AIDS), or when focusing on more demanding concepts such as those encountered in Physics. The absence of studies focusing on aspects of Chemistry has been a consistent feature of review work on small-group discussions, and does not lend itself to very obvious explanation.

Ages of learners in studies

The studies were undertaken with a diversity of age ranges of learners, as summarised in Table 4.4. The ratio (3.2:1) of studies between senior secondary level (ages 16–18) and junior secondary (ages 11–15) is slightly higher than that of studies in the map (4.7:1). Studies judged to be medium-high in overall quality are asterisked (*).

Table 4.4: Ages of learners in studies selected for in-depth review (N = 19, mutually exclusive)

Age range	Number of studies	Study
16–18	6	De Vries *et al.*, 2002 Finkel, 1996 Jiménez-Aleixandre and Pereiro-Muñoz, 2002 Johnson and Stewart, 2002 *Roth and Roychoudhury, 1992 Tao, 2001
13–15	10	*Hogan, 1999a Hogan, 1996b *Jiménez-Aleixandre *et al.*, 2000 *Keys, 1997 Lajoie *et al.*, 2001 Meyer and Woodruff, 1997 *Richmond and Striley, 1996 *Tolmie and Howe, 1993 Tsai, 1999 *Zohar and Nemet, 2002
11–12	3	Kurth *et al.*, 2002 Palinscar *et al.*, 1993 Woodruff and Meyer, 1997

Overall, this section indicates that the studies in the in-depth review are representative of those in the systematic map in terms of reflecting country of study, science subject focus and age of students.

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

Approach to synthesis

This section synthesises the data extracted from the 19 studies, concentrating on the seven studies judged to be of medium-high quality overall. The chief characteristics that these studies have in common is that they have sound methodology and analysis, and the focus is particularly relevant to the review question. Studies of medium quality were also examined, although in less detail. Further details of the evidence which informed the judgements about quality may be found in sections 4.3.3. and 4.3.4.

Section 4.3.1 provides an overview the studies and their aims.

In section 4.3.2, methodological considerations are brought together in order to indicate how judgements were reached about the quality of the studies (Weight of Evidence A).

Section 4.3.3 looks at the research design of the studies in relation to the in-depth review question in order to indicate how judgements were reached about the appropriateness of the study design for the in-depth review question (Weight of Evidence B).

Section 4.3.4 addresses the focus of the studies in relation to the in-depth review question in order to indicate how judgements were reached about the relevance to the in-depth review question (Weight of Evidence C).

The discussion in this section should be read in conjunction with Appendices 2.5 (Enhanced data extraction tool), 2.6 (Indicators for Weight of Evidence) and 4.1 (summary tables).

4.3.1 Overview of the studies

Focus of studies

Hogan (1999a) This study explored students' roles during a long-term collaborative task which required them to master complex sets of cognitive, regulatory and social skills needed for building knowledge, largely from their own and their peers' ideas and observations. The students were in the eighth-grade in the USA, working in eight groups of three, and the discussion focus related to their ideas about solids, liquids and gases.

Jiménez-Aleixandre *et al.* (2000a) This study explored students' capacity to develop and assess arguments during the teaching of a unit on genetics. The students were in the ninth grade in Spain, working in six groups of four students.

Keys (1997) This study explored the reasoning discourse of ninth-grade students in the USA, working in pairs. The report focused on one pair of female students and one pair of male students. Two activities formed the focus for the discussions, one involved making and refining predictions about the contents of a 'black box' and the other involved developing explanations for the reaction between iron and oxygen.

Richmond and Striley (1996) This study explored student talk during four laboratory investigations to help understand the process by which students develop scientific arguments and solve scientific problems. The students were tenth-grade students in the USA, working in six discussion groups of four members each.

Roth and Roychoudhury (1992) This study explored student discourse when compiling concept maps as a means of reviewing and organising knowledge. The students were aged 17–18 and taking either a junior or senior physics course at a college in Canada. They worked in groups of three or four. The detailed data reported focused on the development of ideas about the quantum nature of light and reported on one group only.

Tolmie and Howe (1993) This study explored student discourse during the process of undertaking a computer-based task involving the prediction of the trajectories of falling objects. The students were in schools in Scotland and aged 13–15, and worked in pairs.

Zohar and Nemet (2002) This study explored the outcomes of a unit which explicitly integrated general reasoning patterns into the context of a teaching unit on human genetics. This involved two particular dilemmas which the students were asked to resolve through discussion. The students involved were ninth-grade students from two schools in Israel, working in discussion groups of five to seven members.

Aims of studies

The focus of this review means that the aims of the studies relate to aspects of exploration of the nature of the dialogue which takes place and the ways in which it illustrates students' understanding of evidence. Thus the aims of the studies make reference to terms such as 'conversational dynamics' (Jiménez-Aleixandre *et al.*, 2000), 'reasoning discourse' (Keys, 1997), 'student arguments' (Richmond and Striley, 1996), 'argumentation skills' (Zohar and Nemet, 2002), and 'impact of exchange of opinions on decision-making and learning' (Tolmie and Howe, 1993). In contrast to the other reviews of small-group discussions (Bennett *et al.*, 2004; Hogarth *et al.*, 2004), where studies tended to have a diversity of aims, not all of which related to small-group discussion work, the studies in this in-depth review had a comparatively specific focus. It is recognised that this, in part, arises from the decision to focus on the studies rated as medium-high, but it is argued that this decision lends a very helpful clarity of focus to the review.

4.3.2 Methodological considerations (Weight of Evidence A)

Study designs

All the studies involved exploration of relationships, specifically looking for links between the nature of the small-group discussions and students' understanding of evidence.

Nature of the discussion groups

Groups size varied from pairs to groups of five or six. Table 4.5 shows the size of group employed in each of the medium-high rated studies.

Table 4.5: Size of group (N = 8, not mutually exclusive)

Group size	Number of studies	Study
2	3	Keys, 1997 Tolmie and Howe, 1993 *Zohar and Nemet, 2002
3–4	4	Hogan, 1999a Jiménez-Aleixandre *et al.*, 2000 Richmond and Striley, 1996 Roth and Roychoudhury, 1992
5–6	1	*Zohar and Nemet, 2002

* Zohar and Nemet (2002) used pairs of students which then merged into larger groups of five or six.

Grouping strategy

The most common strategy for creating groups was for the researchers deliberately to create heterogeneous groups (Hogan, 1999a, 1999b; Keys, 1997; Tolmie and Howe, 1993). In some cases, particular care was taken to try to promote argumentation by pairing students with differing characteristics, such as gender (Richmond and Striley, 1996), ability (Richmond and Striley, 1996) and level of understanding (Tolmie and Howe, 1993).

Three studies did not give any specific details of how groups were formed (Jiménez-Aleixandre *et al.*, 2000; Roth and Roychoudhury, 1992; Zohar and Nemet, 2002). In these cases, it seems likely that the groups were formed on the basis of friendship and/or geographic proximity (i.e. students were placed in groups with those sitting nearest them in the class). It would also appear that, by implication, such groups were seen as typical of groups in normal teaching situations.

All the studies involved synchronous small-group discussions – that is, the discussion took place in real time and face to face, rather than in an asynchronous way (e.g. through delayed response and often using computers).

Sample size and sampling method

The nature of the review topic means that studies are likely to involve comparatively small samples.

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, 1999a; Richmond and Striley; 1996; Roth and Roychoudhury, 1992; Tolmie and Howe, 1993). Such convenience sampling is probably realistic for research studies fitting in with practice, and requiring extensive periods of data collection and thus a high degree of co-operation with the class teachers involved.

The sampling strategy varied from study to study. With one exception (Zohar and Nemet, 2002), all the studies were based in one school. In the case of Zohar and Nemet (2002), two schools were used. Some studies sampled students from more than one class: Hogan (1999a) used eight groups of three students each from four classes; Keys (1997) used three pairs, selected from three classes; and Tolmie and Howe (1993) report on 73 students in pairs from different classes, although they do not say how many.

Other studies were limited to groups within one class only. Jiménez-Aleixandre *et al.* (2000) used six groups of four students (although the study reports in detail on only one group of four). Richmond and Striley (1996) also used six groups of four students.

Roth and Roychoudhury (1992) use a sample consisting of 46 and 48 students on a junior level physics course in years 1 and 2 of the study respectively, and 29 and 25 students on the senior level physics course in years 1 and 2 of the study respectively. It is not clear if the students in each year were taught in one or more classes. The data ultimately presented focus on one group of three male students of varying ability.

Zohar and Nemet (2002) had a rather larger sample size drawn from two schools. They had 186 participants overall, assigned to a control group (99 students, five class sets) and an experimental group (87 students, four class sets). Students worked in pairs and then in groups of six.

Two studies justified the sample. Jiménez-Aleixandre *et al.* (2000) purposely selected a class as being midway between student- and teacher-centred. (However, no rationale is given for the selection of the group of four which forms the basis of the detailed analysis.) Keys (1997) selected groups which she felt contained ‘typical students’. In most cases, no justification was given. This may in part be due to the opportunistic nature of the sampling, and in part due to an assumption that the groups generally contained ‘typical’ students.

One characteristic of the work was the use of retrospective sampling – that is, data were gathered on a number of groups, but presented on only a sample of the groups within this, depending on characteristics of the discussion which emerged in the analysis. The studies by Jiménez-Aleixandre *et al.* (2000), and Roth and Roychoudhury (1992) exemplify this approach.

Methods used to collect data

The principal methods of data collection involved obtaining a detailed, usually verbatim, record of student discussions. Thus, virtually all studies used video recording (Hogan, 1999a; Keys, 1997; Richmond and Striley, 1996; Roth and Roychoudhury, 1992) and/or audio recording (Hogan, 1999a; Jiménez-Aleixandre *et al.*, 2000; Zohar and Nemet, 2002). In many cases, this was supported by direct observation (Hogan, 1999a; Jiménez-Aleixandre *et al.*, 2000; Keys, 1997; Tolmie and Howe, 1993; Roth and Roychoudhury, 1992). Hogan, Keys, and Richmond and Striley supported their observation with field notes. Some studies involved extensive video-recording or audio-recoding: for example, Hogan (1999a) taped 73 sessions, and Richmond and Striley (1996) obtained 60 hours of audio recordings and eight hours of video recordings. This amount of data clearly has implications for the fraction that can be presented in a comparatively short report of a study, such as a journal article.

Other sources of data included interviews (Hogan, 1999a; Keys, 1997), student products of tasks, such as the laboratory reports generated in Keys’ study, the concept maps generated in Richmond and Striley’s study, and computer records of predictions made in Roth and Roychoudhury’s study. Students’ views were obtained by self-completion questionnaires or reports in three studies (Keys, 1997; Tolmie and Howe, 1993; Richmond and Striley, 1996). In two studies, measures of student knowledge were obtained (Roth and Roychoudhury, 1992; Zohar and Nemet, 2002). Additionally, Tolmie and Howe collected data via computer-recorded textual interactions and student written predictions of experimental outcomes; Roth and Roychoudhury administered a psychological test; and Zohar and Nemet used student worksheets to collect evidence of argumentation skills.

Overall, the picture gained is one of studies collecting a lot of data in an attempt to get as detailed a picture as possible of students’ dialogue and factors likely to affect their understanding of evidence. In all cases, at least two sources of data were gathered, with three or more sources being common, adding depth and richness to the data, and increasing its trustworthiness. Comparatively little detail was generally given about data-collection tools, although this is often the case in journal papers, and probably a function of restrictions on word length.

Justification for methods used

In general, the studies were characterised by a lack of justification for methods used, or discussion of strengths and weaknesses of methods of data collection. Rather, there appeared to be an implicit acceptance that detailed audio- and video-recordings spoke for themselves as being appropriate methods for collecting data on small-group discussions. This is a reasonable assumption, given the purpose of the research.

Reliability/validity/trustworthiness of data-collection tools

The studies were characterised by minimal detail of any data-collection tools, such as areas explored in questionnaires or interview schedules, or of observation schedules. Indeed, the impression gained is that observation was largely undertaken without schedules being developed. Minimal information was provided on checks on reliability and validity of data-collection tools, with the exception of information provided by Zohar and Nemet (2002) on their multiple-choice instrument assessing biological knowledge. Both these characteristics may be a function of the shortage of space in a research report written for a journal. Alternatively, it may be the case that these tools are seen to provide supporting data for the main data gathered via audio and video recordings.

The Review Group considers that the nature of the studies means that the emphasis is on the trustworthiness of the data collected. The existence of detailed data and multiple data sources in the studies all serve to increase the trustworthiness of the data.

Role of the researchers

In the majority of cases, the researchers were involved in the data collection. The nature of the studies meant that they frequently acted as non-participant observers (e.g. Hogan, 1999a; Jiménez-Aleixandre *et al.*, 2000; Keys, 1997). On the other hand, Roth (Roth and Roychoudhury, 1992) did all the teaching of the classes on which his study is based. In this respect, Roth is the exception to the approaches which appear to be used when investigating small-group discussions.

Comparison/control of independent variable and pre-/post data collection

Given the nature of the studies, it is not surprising that only one of the studies (Zohar and Nemet, 2002) made comparisons between a control group and a group which received some form of intervention related to small-group discussion work. Tolmie and Howe (1993) specifically set up groups where gender was a variable to be explored. However, much of the emphasis of the studies was on describing and interpreting the nature of student discussions and their effects on students' understanding of evidence. Two factors may contribute to the absence of a control group in the studies. Firstly, it is highly likely that those undertaking the research would see no need to design their studies to include a control group in what are largely qualitative and interpretative studies, and this is a reasonable position to adopt. Secondly, the practicalities associated with collecting and analysing extensive in-depth data from a much larger sample in order to make such comparisons would place prohibitive resource constraints on the studies. However, many of the studies make detailed comparisons between groups participating in their studies, as, for example, in the studies of Hogan (1999a), Keys (1997), and Richmond and Striley (1996).

The descriptive and interpretative nature of the research is also likely to be the explanation for the absence of pre- and post data collection relating to the independent variable (the nature of the discussion). Only two studies (Tolmie and Howe, 1993; Zohar and Nemet, 2002) collected pre- and post-data.

Methods used to analyse data

The emphasis of the data analysis is on interpretation of data, much of which came from transcripts of student discourse captured as either video- or audio-recordings. The impression gained from the detailed excerpts included in reports is that the tapes had been transcribed in full, although not all studies stated this explicitly.

Two contrasting overall analysis strategies were apparent in the studies.

The first strategy, adopted in four of the studies, was to develop grounded theory from the data. Hogan (1999a) referred to her analysis as ‘ethnographic interaction analysis’, which she applied to large sections of discourse to identify and interpret patterns of group interactions and roles played by individual students. (These data were supplemented by interview data which were used to develop descriptive profiles of students’ perspectives of learning science.) Roth and Roychoudhury (1992) use what they describe as the techniques used by anthropologists when analysing interactive behaviours. These techniques involve developing categories from the data, which were then used to characterise the interactions between participants and the concept maps they produced. (In this context, ‘participants’ included the teacher as well as the students.) Zohar and Nemet (2002) used qualitative categories developed from earlier research they and others (Resnick *et al.*, 1993; Pontecorvo and Girardet, 1993) had undertaken to score pre- and post-tests of argumentation skills.

The second strategy was to draw on existing work on discourse analysis or discourse analysis classifications, as happened in two of the studies. For example, Jiménez-Aleixandre *et al.* (2000) drew on the work of Bloome *et al.* (1989) to do the initial coding of exchanges between students, and then used Toulmin’s (1958) work on argument to classify the interactions where students were talking about science aspects in the discussion. This process involved breaking down the text into ‘units of analysis’, although no explanation of this term is provided. A third scheme looking at epistemic operations was developed, based on several theoretical classifications and refined on the basis of the data. Keys (1995) drew on elements of a framework developed by Kuhn (1993) to code students’ verbal interactions relating to scientific reasoning and the use of evidence.

Two studies adopted a combination of the above. Richmond and Striley (1996) used three sources of data (audiotapes, videotapes and student notebooks) to identify and categorise concepts with which students were struggling and features of their social interactions. Argumentation ability was classified according to the quality of discussion when students talked about (pre-set) stages of the investigations on which the study focused. Tolmie and Howe (1993) began by developing 13 indices of on-task interactions, based mainly on their videotaped data of students but also drawing on their computer data. Students completed a pencil-and-paper test before and after discussing the tasks, and the indices were used to yield a 'measure of explanation change' by subtracting the post-discussion score from the pre-discussion score. These were calculated for male, female and mixed-gender groups. They then examined patterns of group interaction in more detail, using 'causal analysis' (Blalock, 1972) to identify 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.

Reliability/validity/trustworthiness of data-analysis methods

The nature of the studies meant that statistical analysis of the data was limited. Rather, the emphasis was on the trustworthiness of the methods used.

Hogan (1999a) increased the trustworthiness of her analysis through triangulation of her audio data and field notes and through the presentation of extensive group exchanges to illustrate aspects of student reasoning. She also related her findings to others reported in the literature.

Jiménez-Aleixandre *et al.* (2000) increased the trustworthiness of their analysis by providing several examples of different interpretations of data, drawing on authentic terms used by students. Data are compared across the observation period, and across SDG and whole class interactions. In the discussion, supporting and conflicting evidence from other studies are used to put data analysis in perspective.

Keys (1997) included checks on the reliability of her analysis through independent coding of reasoning strategies of a 10% sample by two researchers, who achieved an inter-coder agreement of 85%. The trustworthiness of her analysis was increased through the triangulation of three sources of data.

Richmond and Striley (1996) checked the reliability of their analysis through independent coding of social roles, with an inter-coder agreement of 100%. Detailed extracts from the conversations were presented. Trustworthiness was increased through the use of multiple data sources, although no corroborating evidence from other studies was presented.

Roth and Roychoudhury (1992) employed a robust procedure for increasing the reliability of their analysis. They report that they adopted a technique used by anthropologists studying interactive behaviours. This involved both researchers watching the videos and reading the transcripts to form tentative descriptions. These were refined, modified or discarded on the basis of further comparisons within sets of data collected. Disagreements were discussed until a consensus was reached, or discarded if no consensus could be reached. Trustworthiness is increased through the use of more than one data source, with dialogue being related to concept maps. The findings are related to a range of other studies.

Tolmie and Howe (1993) drew on an existing scheme to analyse group interactions, which they felt enhanced the reliability of the analysis. They also undertook a reliability check on their observation data, with a 25% sample being independently coded by both researchers with an 81% initial inter-judge agreement. They did not report on steps taken to increase the validity or trustworthiness of their procedures, but drew on multiple data sources. Findings are related to other work in the area.

Zohar and Nemet (2002) both undertook independently the analysis of argumentation skills in their study. Argumentation skills analysis was done by both researchers, and inter-rater reliability scores calculated, although not reported. Some statistical analysis was undertaken, with t-tests carried out on the significance of the use of biological knowledge in the pre- and post-test, and the significance of mean scores on argumentation tests. No details were reported of steps taken to increase the trustworthiness of the analysis, although two data sources were used. Findings are related to other work in the area.

In general, data analysis was characterised positively by the presentation and discussion of rich and detailed data in the form of extracts from students' discourse. However, given that the studies were largely gathering qualitative data, there was a surprising lack of contextual detail. Data also tended to be presented in a rather convergent manner, with few examples of data being presented which might disprove assertions or report on unintended outcomes. A further characteristic of the data analysis was the absence of justification for the study design and for the analysis methods used, and possible limitations to each. The development of grounded theory appeared to be seen as an unproblematic choice for analysis in the majority of cases. However, given that other studies drew on existing models of discourse analysis, some justification for the approach to be adopted appears desirable. There was variation in the detail provided in the studies of methods used to increase the trustworthiness of the analysis, although in all cases the nature of the data and the analysis undertaken appeared to confer a high degree of trustworthiness. Within this, however, it is worth noting that researchers can operate a high degree of selectivity in the examples they choose to present in the reports of their studies.

Additional comments from studies rated as medium in terms of overall quality

The information presented thus far in section 4.3.2 relates to the studies rated as medium-high overall. The medium-rated studies generally tend to confirm the evidence yielded by the medium-high studies. For example, the medium-rated studies all drew on at least two sources of data, with all but one (De Vries *et al.*, 2002) drawing on audio- or video-recordings of group interactions. Grounded theory was the predominant mode of data analysis, with only one study (Jiménez-Aleixandre and Pereiro-Muñoz, 2002) using existing analyses tools. In this case, as with the medium-high-rated study by Jiménez-Aleixandre *et al.* (2000), Toulmin's analysis was used.

4.3.3 Appropriateness of the studies' research design for the in-depth review (weight of evidence B)

Five aspects of each study were examined to reach a view on the appropriateness of the study design for the in-depth review question. These are as follows:

- the nature of the sampling frame

- the actual sample used
- the level of description provided in relation to small-group discussions
- the trustworthiness of the data collection
- the trustworthiness of the data analysis

The nature of the sampling frame

In the majority of cases, the issue of sampling frame did not arise. Rather, the studies made use of opportunistic samples which were available to the researchers through personal contacts. However, Hogan (1999a), Tolmie and Howe (1993), and Zohar and Nemet (2002) did provide some details of the sampling frame in terms of school context(s) and teachers involved.

The actual sample used

A characteristic of many of the studies was the limited information provided to justify the selection of the sample used. Only Keys (1997) provided information in any detail on the characteristics of the group members and the reasons for their selection. It may be that the selection of groups used was not considered to be an issue by the researchers, in the same way that the issue of sampling frame did not arise. However, the scarcity of information about the samples used and the reasons for their selection does seem a surprising omission from qualitative studies.

Another aspect of relevance in relation to the actual sample used concerned the use to which it was put in terms of making comparisons. Five of the studies (Hogan, 1999a; Keys, 1997; Richmond and Striley, 1996; Tolmie and Howe, 1993; and Zohar and Nemet, 2002) made comparisons between groups. For example, Hogan made comparisons on the basis of pre-determined surface/deep reasoning ability, and of findings on individuals' roles, Keys made comparisons on the basis of differences in students' levels of conceptual understanding, and Tolmie and Howe looked at gender differences. Zohar and Nemet employed an experimental design, making comparisons between groups on the basis of whether or not they had received an intervention package on developing argumentation skills. Roth and Roychoudhury (1992) made comparisons within their sample group only, and Jiménez-Aleixandre *et al.* (2000) made only minimal comparison within groups or between groups. It is therefore apparent that the notion of making some form of comparison between students is important to the majority of the researchers.

The level of description provided in relation to small-group discussions

A very important aspect contributing to judgements in relation to weight of evidence B concerned the level of description provided of the context of the small-group discussions. This was a strong feature of all the studies, most particularly those of Hogan (1999a); Jiménez-Aleixandre *et al.*, 2000; Keys, 1997; and Roth and Roychoudhury, 1992.

The trustworthiness of the data collection

The studies were characterised by high levels of trustworthiness in relation to the methods of data collection. Most pertinently in the context of the review question, discussions were normally audio- and/or videotaped and transcribed verbatim, as evidenced by detailed extracts included in the reports (with the exception of Tolmie and Howe (1993), who did not include any verbatim extracts).

The trustworthiness of the data analysis

The studies were also characterised by high levels of trustworthiness in relation to the data analysis, with the majority incorporating sound strategies to maximise such trustworthiness. These included one or more of double coding of responses (with good inter-rater agreement), use of existing classifications, use of multiple data sources, and relating the findings of the study to other relevant literature.

Additional comments from studies rated as medium in terms of overall quality

There was no evidence from the medium-rated studies which contradicts that from the medium-high-rated studies in terms of either approach to design or analysis. Again, studies tended to be characterised by multiple data sources, with the use of audio- and/or video recordings proving a major component of the data. In keeping with this, data analysis very often drew on detailed extracts from transcripts of recordings. As with the medium-high-rated studies, the majority of the medium-rated studies were characterised by the provision of very limited information on the reasons for selecting the sample, although there were some exceptions to this (Finkel, 1996; Hogan 1999b).

There was no single characteristic which distinguished the medium-rated studies from the medium-high-rated studies. Rather, they tended to provide less detail in one or more of the five categories used to reach judgements. Thus, for example, Finkel's (1996) study made little use of quotations in the discussion and analysis; Hogan's (1999b) study lacked checks on the reliability of the analysis; Kurth's (2002) failure to articulate research questions clearly created difficulties in assessing aspects of the analysis; and Woodruff and Meyer (1997) provided almost no contextual detail.

4.3.4 Relevance of the studies' focus for the in-depth review (weight of evidence C)

The following five features of the study designs were selected to establish the appropriateness of their focus for the in-depth review question:

- the focus of the intervention
- the focus of the study
- the measures employed to test the nature of the small-group discussions
- the breadth of data reported

- the representativeness of the situation in which the studies were conducted in relation to normal classroom settings

The focus of the intervention

All the medium-high-rated studies focused directly on aspects of students' understanding of evidence, although different terms were sometimes used to describe this characteristic. For example, Hogan (1999a) explored what she termed 'sense-making model construction'; Richmond and Striley (1996), and Zohar and Nemet (2002) referred to 'argumentation skills'; and Tolmie and Howe (1993) made reference to 'self-generated evidence'.

The focus of the study

With one exception, the nature of the discussion which took place between students working in small groups was an explicit variable explored. The exception was Hogan (1999a), who, whilst looking at aspects of the discussion, set this in the context of the roles played by participants.

The measures employed to test the nature of the small-group discussions

All the studies employed sound measures to test the nature of the small-group discussion, either by developing grounded theory from a solid evidence base, or by using existing analysis tools to analyse student dialogue. As reported earlier, most of the studies drew on detailed transcripts of audio- and/or video recordings to generate their data, supported by observation and field notes in some cases.

The breadth of data reported

All the studies reported a wide breadth of data on aspects of the student discussion, although the ways in which data were reported differed according to the strategy adopted for the analysis. In one case (Roth and Roychoudhury, 1992), the breadth of data reported was limited in that it drew on only one group of three students.

The representativeness of the situation in which the studies were conducted in relation to normal classroom settings

With the exception of one study, all the studies were conducted in normal classroom settings, making the data gathered highly representative of the sorts of situations teachers might encounter in their own classes if they were using small-group discussions. The slight exception was Tolmie and Howe's (1993) study, in the sense that their interest in gender aspects resulted in them applying rather more criteria to the constitution of their groups (gender, responses to predict and explain task) than would be likely to happen in normal teaching situations. Additionally, Roth and Roychoudhury (1992) conducted their study in an independent college, which makes their data less representative of the student population as a whole.

Additional comments from studies rated as medium in terms of overall quality

There was no evidence from the medium-rated studies which would contradict that from the medium-high-rated studies. There were two principal factors which contributed to the medium-rated studies being judged to provide a less solid evidence base for the review. The first of these concerned the use of an atypical sample. De Vries's (2002) sample consisted of volunteers participating in an after-school activity, and Jiménez-Aleixandre and Pereiro-Muñoz (2002) gathered their data in an evening class which contained students in a wide age-range, including adult learners. The second factor was the extent to which the study reported directly on the nature of the discussion. In three cases (Finkel, 1996; Hogan, 1999b; Lajoie, 2001), the study reported only indirectly on the discussion. In the case of the other two studies, the limitations concerned either the focus (Kurth *et al.*, 2002, did not have understanding of evidence as their main focus), or the discussion reported (Woodruff and Meyer, 1997, present only a very narrow range of discussion).

4.4 Synthesis of evidence

Although the seven studies rated as medium high share a number of features in common in relation to the study design and data analysis, there is considerable diversity in the research questions addressed and the topic used to promote small-group discussion. For example, Jiménez-Aleixandre *et al.* (2000) had a clear focus on looking at patterns in the development of arguments, whereas Roth and Roychoudhury (1992) were interested in documenting discussion in relation to students' development of concept maps. Only two studies had some overlap in discussion topic: Jiménez-Aleixandre *et al.* (2000) explored students' capacity to develop and assess arguments during the teaching of a unit on genetics; and Zohar and Nemet (2002) explored the effects of a unit which explicitly integrated general reasoning patterns into the context of a teaching unit on human genetics. The topic of genetics is one which would seem likely to lend itself well to discussion tasks.

Hogan (1999a) used long-term collaborative tasks which required students to develop and refine their ideas about solids, liquids and gases. Keys' (1997) study focused on two short activities, one involving making and refining predictions about the contents of a 'black box' and the other involved developing explanations for the reaction between iron and oxygen. Richmond and Striley (1996) explored student discourse during four laboratory investigations. Roth and Roychoudhury (1992) explored student discourse when compiling concept maps to develop ideas about the quantum nature of light. Tolmie and Howe's (1993) study looked at discourse when students were undertaking a computer-based task involving the prediction of the trajectories of falling objects

One outcome of this diversity in research questions and discussion topics is that some of the findings do not lend themselves easily to fitting into any particular pattern, as they are very specific to a particular study. However, some common findings have emerged across studies.

This section therefore presents both study-specific findings where they offer insights into the review question, and seeks to establish common cross-study themes. The study findings are considered under five headings:

- group structure and interaction dynamics between group members
- ways in which meaning was reached

- features of discussion topic
- claims made for effects of small-group discussion work on students' understanding of evidence
- need for specific instruction about making arguments

Finally, other findings of interest from medium-rated studies are considered.

Group structure and interaction dynamics between group members

Some studies report findings which focused on group structure and interaction dynamics between group members. The principal evidence came from four studies: Hogan (1999a), Keys (1997), Tolmie and Howe (1993), and Richmond and Striley (1996), with both Hogan and Richmond and Striley looking in detail at the roles of group members.

Hogan (1999a) identifies eight sociocognitive roles in group reasoning processes which took place in small-group discussions, all of which were persistent over time. Four of these were positive (promoter of reflection, contributor of content knowledge, creative model builder, mediator of social interaction and ideas) and four were negative (promoter of distraction, of acrimony, of simple task completion, reticent participant). She indicates that the roles of leader and helper, identified in other literature, did not emerge in the context of her work. She established that at least one group member had to act in a way which promoted reflection in the group for deep (as opposed to surface) reasoning to take place. Hogan concludes that assigning managerial or prosocial roles to students (e.g. reflector, regulator, questioner, explainer) as suggested in collaborative learning theory was likely to be counter-productive for ill-structured intellectual tasks. She also concludes that friendship groups were more effective in developing meaning-making abilities and higher-order thinking in small groups.

Richmond and Striley (1996) established that progress was dependent on group dynamics, and particularly the style of the group leader. Where the leader adopted an inclusive style, this allowed substantial engagement in the discussion by a number of participants, and increased the quality of the discussion. This, in turn, permitted most members to succeed in connecting new knowledge to the larger intellectual picture. Persuasive leadership allowed high engagement of the leader, but engagement of other members was limited to procedural matters rather than to discussion which improved understanding. Thus only the leader developed connections between ideas. Alienating leadership generated a lot of off-task talk and engagement was generally low, with little concern for making connections between ideas. Whilst the quality of argument was high in inclusive and persuasive groups (co-constructed in the first case), groups with alienating leadership had fragile arguments and had trouble substantiating their claims under scrutiny. In conclusion, Richmond and Striley argued that their three goals (engagement, placing new knowledge in intellectual context, construction of argument) were best supported by requiring distributed responsibility during group presentations, by making each individual in a group produce their own report, and by fostering a style of inclusive leadership and equitable classroom participation.

Tolmie and Howe (1993) focused on aspects of the gender composition of groups. They found improved explanations to be independent of group composition (male, female, mixed), with the biggest improvements being noted when groups contained members with a high degree of dissimilarity in their initial predictions and explanations. Whilst improvements were independent of gender composition, there were clear differences in interactional styles and manner of progress. Male pairs were more willing to exchange ideas and engage in explicit co-ordination of ideas and evidence, although they tended not to generalise ideas across problems. Female pairs tended to avoid conflict and look for co-ordination of ideas relevant to different problems. Mixed pairs interacted in a more constrained way, although again avoiding conflict. Tolmie and Howe (1993) suggest that the process of opinion exchange was central to the development of ideas, and that both male pairs and female pairs demonstrate qualities one would want to see in the development of arguments. However, they do not see this as a reason for promoting the use of mixed gender pairs, as their study suggests the best of the all-male and all-female pair interactions were lost in mixed pairs. They indicate that they hope to engage in further work looking at ways of structuring the discussion topic to introduce specific aspects aimed at helping male pairs develop the qualities female pairs brought to the task, and vice versa.

Like Tolmie and Howe, Keys (1997) also found that male pairs enjoyed the challenge of testing their ideas and of debating the correctness of various scientific ideas, whereas female pairs on the other hand were much less likely to engage in evaluating and critiquing one another's ideas. Thus, the factor of gender in group composition had a noticeable impact on the ways in which debate developed. Keys also notes that the greatest improvement in reasoning discourse occurred in pairs who were initially reluctant to discuss the meaning of scientific concepts.

Ways in which meaning was reached

Three of the studies reported in detail on the ways in which meanings were negotiated and agreed: Jiménez-Aleixandre *et al.* (2000b), Keys (1997), and Roth and Roychoudhury (1992).

Jiménez-Aleixandre *et al.* (2000b) established that a large proportion of discourse statements relate to what they termed 'doing the lesson' (interaction referring to the rules of the task, or to perceived features of science classrooms), with rather fewer relating to the science involved. Later in the discussion, statements indicating 'doing science' increase. More specifically, and drawing on Toulmin's framework (Toulmin, 1968), Jiménez-Aleixandre *et al.* (2000b) report that 35% of arguments in small-group discussions were claims, 20% were warrants, 10% drew on data, and 5% are backings. The other two categories, rebuttals and qualifiers, only occurred in whole-class plenary discussion. Jiménez-Aleixandre *et al.* (2000b) also noted that arguments were frequently developed by a subset within the group and, although agreement was generally reached, this was often for social reasons, with deviating personal opinions still persisting.

The findings reported by Jiménez-Aleixandre *et al.* (2000a) are echoed by Roth and Roychoudhury (1992), as they also found that students tended not to engage very often in processes which fostered meaning; rather they would reach agreement on the basis of finding something agreeable to all group members. Within this, students might well form strategic alliances in support of a position, with positions seen as having more weight if a student were known to have a special interest in the area. Agreements were often reached by one or more group members exerting authority, on the basis of majority rule or by acceptance of a common lower order of agreement. However, agreement was not always based on a common understanding.

More positively, Roth and Roychoudhury (1992) report a range of interactions within groups. Discussions often involved positions being stated, contested and views either accepted or temporarily or permanently rejected, with temporarily rejected positions sometimes becoming accepted as positions finally stabilised into shared meaning as the concept maps were constructed.

Keys (1997) established three characteristics of reasoning in discussions: recognising that prior ideas (models) may be incorrect; evaluating new observations for consistency with current ideas and using evidence to modify ideas; and co-ordinating all mutually consistent knowledge propositions into a coherent model. She further suggests that scientific reasoning could be identified by 11 skills clustered into four categories of reasoning skills: (a) assessing prior models (posing predictions; evaluating predictions; explaining/justifying predictions): (b) generating new models (evaluating observations; identifying patterns; drawing conclusions; formulating models); (c) extending models (inferring; comparing/contrasting); and (d) for support (discussing concept meaning; identifying relevant information).

Features of discussion topic

This review did not set out to explore the nature of the discussion topic in detail, as this has been the focus of another review (Hogarth *et al.* 2004). However, some of the studies did make specific observations on the relation of the discussion topic to the nature of the discussion. Some of these related to the need for the discussion topic to provide opportunities for both internal and external debate/conflict, if students' understanding of evidence is to be improved. For example, Tolmie and Howe (1993) required students to make individual predictions, then engage in a task which required a joint prediction (internal debate/conflict), and finally to compare this with an actual situation to reach an explanation of any discrepancies (external debate/conflict). Whilst other studies did not comment on this specifically, some of the accounts indicate that internal and external debate were built into student tasks, as was the case for Zohar and Nemet (2002).

Two studies offered comments on the nature of the data provided to students for the discussion. Jiménez-Aleixandre *et al.* (2000b) indicate that hypothetical, unquestionable data (provided by the teacher) generates different patterns of argument compared with empirical, uncertain data, perhaps gathered by students themselves, whilst Roth and Roychoudhury (1992) advocate the use of a fixed set of concepts to delimit the content of the discourse.

Claims made for effects of small-group discussion work on students' understanding of evidence

Whilst the primary focus of the studies included in this in-depth review was on the nature of the discussion and the ways in which teachers used small-group discussions, all the researchers also drew some conclusions about the effects of the discussions. Indeed, it would be highly surprising if they had not done so, as activities are rarely introduced into lessons without some purpose, and the purpose often relates to developing understanding.

All the researchers were of the view that small-group discussions assisted students in developing arguments and making use of evidence in science. However, a note of caution needs to be sounded, as the focus of the small group discussion work was very varied. Set against this, the in-depth review has focused on studies rated as medium-high in terms of their overall quality, which suggests that confidence can be placed in the findings, albeit in a somewhat limited way. The following examples illustrate the variety of circumstances in which small-group discussion work was found to enhance students' use of evidence.

Small-group discussions were seen to enhance students' understanding of evidence in investigations, as reported by Richmond and Striley (1996), who demonstrated that students showed both increasing levels of sophistication and increased use of subject knowledge in the arguments they developed during the discussions.

Roth and Roychoudhury's (1992) study used the development of concept maps as the vehicle for promoting small-group discussions. They established that group discourse over the construction of a concept map provided a vehicle for negotiation of meaning and understanding of concepts and their relationships, thus providing a structure through which students were able to learn the language patterns of science and use these to construct scientific knowledge. Roth and Roychoudhury further concluded that mapping concepts as a group activity might well be more important than the concept map itself.

Tolmie and Howe (1993) explored student discourse while undertaking a computer-based task involving the prediction of the trajectories of falling objects. They report that their intervention resulted in significant improvements in students' explanations, and suggest that group-orientated software which encourages joint decisions would be worth developing in the teaching of physics. They also made recommendations for ways in which it could be adapted for use with female and mixed pairs of students who emerged as weaker than male pairs at making predictions.

Zohar and Nemet's (2002) study explored the outcomes of a unit which explicitly integrated general reasoning patterns into the context of a teaching unit on human genetics. They report 'dramatic changes' in the quality of student arguments, with students justifying claims more frequently, and using more sophisticated ideas, including specific biological knowledge, to make more complex arguments, which included an increased number of justifications for the conclusions they reached. This finding is very similar to that of Richmond and Striley (1996). Zohar and Nemet (2002) also report a more general improved understanding of the biological concepts associated with the genetics module. This led them to recommend that reasoning about dilemmas should be integrated into other science topics.

Two studies (Roth and Roychoudhury, 1992; Zohar and Nemet, 2002) commented specifically on the length of time students were able to sustain small-group discussion work, finding that typical discussions would occupy most of a lesson.

Need for specific instruction about making arguments

The most significant cross-study message to emerge from the review concerns the need for specific instruction about making arguments and using evidence. Five of the seven studies rated as medium-high make specific mention of aspects relating to this, and a further three of the medium-rated studies also make reference to this in their findings.

Of the medium-high studies, three (Hogan, 1999a; Richmond and Striley, 1996; Zohar and Nemet, 2002) all suggest some training in skills in handling group discussions is vital, and two (Jiménez-Aleixandre *et al.*, 2000a and Roth and Roychoudhury, 1992) recommend coaching in argumentation skills.

Hogan (1999a) argues that guiding students towards taking constructive roles in discussions may be achieved through metacognitive training – that is, knowledge about the nature of collaborative learning, effective group learning strategies, and awareness of what constitutes progress.

Richmond and Striley (1996) indicate that productive learning is unlikely to take place on a large scale through the use of small-group discussions until students acquire the skills associated with inclusive leadership and are thus able to foster a climate of equitable participation.

Zohar and Nemet's (2002) study did involve incorporating explicit instruction about augmentation into their intervention. One introductory lesson involved arguments being defined and their structure explained. Characteristics of good arguments were also given so that students were made aware of the need to identify and include true, reliable and multiple justifications and also refer to alternative explanations and rebut them. Students then practised the principles through several concrete examples. Zohar and Nemet conclude that 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.

Jiménez-Aleixandre *et al.* (2000a) suggest their work indicates that, if the ability to develop arguments is set as a learning goal in science, it will not happen during normal instruction. Rather, they argue for a need to provide specific, inquiry-focused tasks where help is given to students to develop their understanding through the construction of arguments.

Roth and Roychoudhury (1992) report that students frequently struggled with language, often making short utterances, and appeared to find it difficult to clarify their understanding through explanations, justifications and elaborations. This led them to conclude that a major outcome of their study was the recognition of the need to help students to argue and to use evidence to support a proposition.

Other findings of interest from medium-rated studies

The findings of the medium-rated studies are in keeping with those of the medium-high-rated studies, although some offer additional insights in the form of notes of caution.

As was the case with the medium-high-rated studies, all the authors of the medium-rated studies report positive outcomes in relation to students' use of evidence as a result of engaging in small-group discussion work. For example, Woodruff and Meyer conclude that small-group discussion helped students refine their explanations and promote their understanding of ideas.

Three studies (Finkel, 1996; Hogan, 1999b; and Kurth *et al.*, 2002) advocate the need for explicit instruction on the nature of arguments and the promotion of a group atmosphere so that all students are able to contribute. However, Hogan also noted that such knowledge did not always translate into improved collaborative reasoning behaviours, nor the deeper processing of ideas and information that would have been manifest in an enhanced ability to apply conceptual knowledge. In contrast to Hogan, who suggested that roles should emerge naturally in groups discussing complex tasks, Kurth *et al.* advocated assigning particular roles to pupils in groups as a means of achieving this. This contrast may well be related to the focused tasks used by Kurth *et al.* However, both Kurth *et al.* and Hogan strongly support the notion of promoting 'inclusive leadership' in groups, a factor they recognise as being very difficult.

De Vries *et al.* (2002) confirm that internal and external debate/conflict were key aspects in developing argument. However, they also note that simply putting students holding different ideas into groups was not a sufficient condition; the group climate needed to be such that students also wanted to discuss their ideas.

4.5 Reflections on the use of the enhanced data-extraction tool

This section reports on the use of the enhanced data-extraction tool (see Appendix 2.5). The reasons for developing this tool, and the process by which it was developed have been described in section 2.3.2.

The enhanced data-extraction tool asked for specific details of relevance to the quality of qualitative studies to be entered into EPPI-Reviewer. These details related to the design of the study, important features of the data collection, important features of the analysis, and ethical considerations.

Overall, the tool was found to be very helpful in systematically identifying and recording details of studies which were not addressed in the standard data-extraction tool. In addition to the information yielded, the enhanced data-extraction tool also served to identify areas where qualitative studies provided good and appropriate detail, and areas where more detail would have been helpful.

Study design: A number of the studies which were data-extracted provided little detail about the sample and the way in which it was identified. Equally, relatively little information was routinely provided about strengths and weaknesses of data sources and methods. Both these aspects provide important information about a study, and were important omissions. Details of the status and roles of the researcher were also often omitted, although this aspect is not unique to reports of qualitative studies.

Important features of data collection: In general, sufficient details were provided about methods of data collection, although examples of instruments, such as observation protocols/schedules or interview schedules (or illustrative extracts from such instruments) were not included. In the case of this review, this absence may well be related to the review focus and principal data-collection technique: that is, there are no examples of 'data-collection tools' which could be included when data are gathered by audio- and/or video-recordings. It may also be a function of lack of space in a journal paper.

Although there was generally good evidence in the reports of steps taken to increase the trustworthiness of the data collected, these were seldom articulated directly in the reports.

In general, the studies were characterised by an acceptance of approaches and methods of data collection which were not seen to require justification.

Important features of analysis A number of very positive features emerged in the analysis. The studies typically provided good detail about descriptive analytic categories used in the analysis. Sound procedures were often adopted, and reported, for increasing the reliability and trustworthiness of the analysis. Findings were normally related to other relevant work in the area. Less positively, data tended to be presented in a convergent manner, with few examples presented which looked at exceptions to emerging patterns, and possible explanations for these. Further, there did not seem to be any significant attempt to look for and comment on unintended or unanticipated consequences arising from undertaking the studies.

Ethical considerations The studies were characterised by an absence of detail on matters relating to consent and confidentiality. Whilst this is not limited to qualitative studies, the emphasis of such studies on looking in detail at a comparatively limited number of people does make this a particularly surprising omission.

4.6 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 2.6) 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.

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 as follows:

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

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

What is the nature of small-group discussions aimed at improving students' understanding of evidence in science?

5.1.2 Mapping of all included studies

There were 94 studies 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 USA, 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; less relates to students' attitudes.
- A diversity of measures was used to assess effects on understanding and attitude.
- Very little research has been done on small-group discussions in relation to the teaching of chemistry.
- Typical small-group discussions 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 the case study.
- There were 28 studies with experimental designs, of which 12 were randomised controlled trials (RCTs).

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

5.1.3 Nature of studies selected for in-depth review

Nineteen studies met the inclusion criteria for the in-depth review set out in section 2.3.1. Table 5.1 summarises the overall weights of evidence assigned to each of these studies.

Table 5.1 Overall weights of evidence (WoE) assigned to the included studies (N = 19)

Medium-high (MH)	Medium (M)	Medium-low (ML)
Hogan, 1999a Jiménez-Aleixandre *et al.*, 2000 Keys, 1997 Richmond and Striley, 1996 Roth and Roychoudhury, 1992 Tolmie and Howe, 1993 Zohar and Nemet, 2002	De Vries *et al.*, 2002 Finkel, 1996 Hogan, 1999b Jiménez-Aleixandre and Pereiro-Muñoz, 2002 Kurth *et al.*, 2002 Lajoie *et al.*, 2001 Woodruff and Meyer, 1997	Johnson and Stewart, 2002 Meyer and Woodruff, 1997 Palinscar *et al.*, 1993 Tao, 2001 Tsai, 1999
(7)	(7)	(5)

The review synthesis concentrated on the seven studies judged to be of medium-high weight of evidence, as they provide the strongest evidence base on which to make recommendations.

5.1.4 Synthesis of findings from studies in in-depth review

Nineteen studies were included in the in-depth review, which focused on the nature of small-group discussion work aimed at improving students' understanding of evidence.

The consolidated evidence from the review draws primarily on the findings from studies weighted as medium-high and, to a lesser extent, as medium in terms of their overall quality.

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

Findings on nature of small-group discussions

Although there is considerable variety in the detailed research questions and discussion topics used to promote small-group discussion, there is a high degree of consistency in the findings and conclusions. In general, students often struggle to formulate and express coherent arguments during small-group discussions, and demonstrate a relatively low level of engagement with tasks. The review presents *very strong evidence of the need for teachers and students to be given explicit teaching in the skills associated with the development of arguments and the characteristics associated with effective group discussions*. Five of the seven highest quality studies in the review make this recommendation. There is also *good evidence* to confirm the findings of other reviews on small-group discussions (Bennett *et al.*, 2004; Hogarth *et al.*, 2004) on the desirability of the stimulus used to promote discussion involving both internal and external conflict, i.e. where a diversity of views and/or understanding are represented within a group (internal conflict) and where an external stimulus presents a group with conflicting views (external conflict).

There is *good evidence* on group structure. Not all studies addressed this aspect, but where advice is offered, it tends to indicate that groups should be specifically constituted such that differing views are represented. There is also evidence to suggest that assigning managerial roles to students (e.g. reflector, regulator, questioner, explainer) as suggested in collaborative learning theory is likely to be counter-productive for poorly-structured tasks. Some evidence is also presented which suggests single-sex groups may function better than mixed-sex groups, although overall development of understanding is not affected by group composition. Group leaders also emerge as having a crucial role: those that were able to adopt an inclusive style, and one which promoted reflection, were the most successful in achieving substantial engagement with the task. An alienating leadership style generates a lot of off-task talk and low levels of engagement.

The review presents some evidence that small-group discussion work does improve students' understanding and use of evidence. Whilst this was not the main focus of the review, all the studies included present some evidence in this area, as improvement in use of evidence was one of the reasons for using small-group discussions. The effects of small-group discussions on students' understanding of evidence has been explored in more detail in other reviews (Bennett *et al.*, 2004; Hogarth *et al.*, 2004).

Findings on research strategies adopted to explore aspects of small-group discussion work

A number of similarities emerged in the approaches adopted in the studies. They tended to make use of opportunistic samples, drawing on the researchers' personal contacts. Experimental designs are not often used, although studies often made comparisons between discussion groups in the same class or within a discussion group. Data collection methods typically involve audio- and/or video-recordings, with analysis and reporting drawing heavily on extracts from recorded dialogue. Whilst approaches to gathering data are seldom justified in any detail by the authors, sound procedures appear to be introduced to check the reliability of the data analysis and to present the findings in a way which makes them trustworthy.

A key difference which has emerged concerns the two contrasting approaches to data analysis, with some studies developing grounded theory from the data, and others drawing on existing models to structure their analysis.

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 and from the dramatic rise in published papers since 1992, as demonstrated in the map for this review (see Table 3.1).
- The review has served to establish that there is consistency in the research approaches that those working in the area feel are appropriate to researching practice related to the use of small-group discussions. Such approaches draw extensively on qualitative data in the form of audio- and/or videotapes of dialogue during discussions, interview data and students' written responses.
- The review has deliberately focused on synthesising the evidence from the medium-high-rated studies.
- 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 one principal limitation. Although the studies in the in-depth review shared a number of similar characteristics at the broad level, there were substantial differences at the detailed level. For example, there was considerable variety in the specific research questions, the topics used for the discussion tasks, and in the use and interpretation of key terms relating to *evidence* and *understanding of evidence*. However, the effect of this limitation was minimised by focusing the in-depth review on studies of medium-high quality. (No studies were rated as high quality.) Whilst it is felt that the review has been strengthened by this limited focus, it could be useful to test the effects of not including other studies through a sensitivity analysis.

5.3 Implications

5.3.1 Policy

Current policy strongly advocates the use of small-group discussion work. Whilst the main focus of this review was to look at the nature of small-group discussions and establish *how* they were being used in science lessons, it also yielded evidence of some potential benefits in terms of helping students develop their skills in formulating arguments. Hence the review does indicate that there could be benefits in pursuing such a policy, but one which recognises that there is more research which needs to be undertaken (see section 5.3.3) before more definite conclusions can be drawn. *However, it is clear from the review that small-group discussion work needs to be supported by the provision of guidance to teachers and students on the development of the skills necessary to make such work effective*. Thus, some form of professional development training for teachers would appear to be highly desirable to provide them with guidance on how to maximise the effectiveness of small-group discussions.

5.3.2 Practice

The review indicates that appropriately-structured small-group discussion work may well provide an effective vehicle for assisting students in the development of ideas about using evidence and constructing well-supported arguments. Thus teachers should be encouraged to incorporate such discussions into their teaching, provided that appropriate support is offered (see sections 5.3.1 and 5.3.3).

5.3.3 Research

Secondary research

The review indicates that the most useful form of secondary research which could be pursued would be to look at methods used to analyse student discourse to establish similarities and differences in existing frameworks and frameworks emerging from grounded theory.

Primary research

One particularly strong feature which has emerged from the work undertaken for this review and the others on small-group discussion (Bennett *et al.* 2004; Hogarth *et al.* 2004) 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, linked to a coherent analysis framework drawing on the findings of the secondary research proposed above.

6. REFERENCES

6.1 Studies included in map and synthesis

The 94 studies included in the systematic map were reported in 119 papers. For the purpose of the map and synthesis, one paper was selected as the lead paper for each study. Subsidiary papers are marked with an asterisk (*).

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

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

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

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

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

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

Chang 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.

Goldman SV (1996) Mediating microworlds: collaboration on high school science activities. In: Koschmann T (ed.) *CSCL: Theory and Practice of an Emerging Paradigm. Computers, Cognition and Work*. New Jersey: Lawrence Erlbaum Associates Inc., pages 45–81.

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.

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

*Jiménez-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.

Jiménez-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.

Jiménez-Aleixandre MP, Rodriguez AB, Duschl RA (2000a) ‘Doing the lesson’ or ‘doing science’: argument in high school genetics. *Science and Education* **84:** 757–92.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Roschelle J (1996) Learning by collaborating: convergent conceptual change. In Koschmann T (ed), *CSCL: Theory and Practice of an Emerging Paradigm. Computers, Cognition and Work*. New Jersey: Lawrence Erlbaum Associates Inc, pages 209–248.

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

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

Roth 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.

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

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

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

Tao 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., pages 229–257.

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

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, University of London.

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

Blalock H (1972) *Causal models in the social sciences*. London: Macmillan.

Bloome D, Puro P, Theodorou E (1989) Procedural display and classroom lessons. *Curriculum Inquiry* **19:** 265–291.

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.

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

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

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

Kuhn D (1993) Science as argument: implications for teaching and learning scientific thinking. *Science Education* **77:** 319–337.

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.

Pontecorvo C, Girardet H (1993) Arguing and reasoning in understanding historical topics. *Cognition and Instruction* **11:** 365–395.

Resnick L, Salmon L, Zeitz C, Wathen S, Holowchak M (1993) Reasoning in conversation. *Cognition and Instruction* **11:** 347–364.

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

Spencer L, Ritchie J, Lewis J, and Dillon L (2003) *Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence*. London: The Strategy Unit (also available at http://www.policyhub.gov.uk/)

Toulmin S (1958) *The Uses of Argument*. New York: Cambridge University Press.

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	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 asynchronous group discussion (e.g. computer mediated)

3. **Not about small-groups**

   – two to six participants

4. **Not on substantive and explicit discussion tasks**

   – explicit discussion tasks taking more than two minutes

5. **If only about effects of group discussions, not about the effect on students' understanding or attitude**

   – understanding includes understanding of science concepts and ideas about science

   – attitude includes attitude to science and to science education

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

   – Out of school can be included.

EXCLUSION ON STUDY TYPE

7. (a) Editorials, commentaries, book reviews or position papers
   (b) Policy documents, syllabuses, frameworks or specifications
   (c) Resources

(d) Bibliography
(e) Theoretical (non-empirical) paper
(f) Methodology paper

EXCLUSION ON SETTING IN WHICH STUDY WAS CARRIED OUT

8. Not published in English

9. Not published in the period 1980–2003

Appendix 2.2: Search strategy for electronic databases

Subject

Small-group discussions in science teaching

Population

Students aged 11 to 18

Limits

English language

1980 to 2003

2.2.1 Educational Resources Information Center (ERIC)

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

1 exp cooperative learning/
2 "ARGUMENTATION".mp.
3 exp discourse analysis/ or exp persuasive discourse/
4 exp discussion/ or exp "discussion (teaching technique)"/ or exp discussion groups/ or exp group discussion/
5 1 or 2 or 3 or 4
6 5 and (science or biology or chemistry or physics or earth science).mp. [mp=abstract, title, headings word, identifiers, full text]
7 limit 6 to (english language and (elementary secondary education or elementary education or intermediate grades or secondary education or middle schools or junior high schools or high schools or high school equivalency programs or postsecondary education or two year colleges) and (books or conference proceedings or dissertations or "evaluative or feasibility reports" or general reports or journal articles or project descriptions or "research or technical reports" or "speeches or conference papers")) and yr=1980–2002

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

1 'cooperative learning' or 'small group instruction' or 'learning strategies' or 'group discussion'
2 'argumentation' or 'verbal communication' or 'discourse analysis'
3 'persuasive discourse' or 'persuasive strategies'
4 'discussion groups' or 'discussion (teaching technique)' or 'discussion'

5 1 or 2 or 3 or 4
6 5 and (‘science’ or ‘biology’ or ‘chemistry’ or ‘physics’ or ‘earth science’)
7 limit 6 to (‘english language’) and (‘secondary education’ or ‘elementary education’ or ‘intermediate grades’ or ‘middle schools’ or ‘junior high schools’ or ‘high schools’ or ‘high school equivalency programs’ or ‘postsecondary education’ or ‘two year colleges’) and (‘books’ or ‘collected works – proceedings’ or ‘dissertations/theses’ or ‘reports – research’ or ‘journal articles’ or ‘speeches/meeting papers’) and yr=2003

2.2.2 British Education Index (BEI)

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

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

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

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

2.2.3 PsycINFO

PsycINFO was searched on 10 April 2003, using the WEBSPIRS interface and 537 records were retrieved. The search of PsycINFO was not updated as experience in the previous review indicated that no papers, other than those already found through ERIC and SSCI, were identified.

1 (cooperative-learning or cooperation or cooperation- or cooperative) in MJ,MN,AG,PO,KC
2 (argument or argumentation) in MJ,MN,AG,PO,KC
3 (discourse-analysis or discourse-processes or discourses) in MJ,MN,AG,PO,KC
4 (discussion-group or group-decision-making or group-discussion or group-dynamics or group-decision-and-negotiation) in MJ,MN,AG,PO,KC
5 1 or 2 or 3 or 4
6 5 and (education* or school* or college or student* or pupil* or learner*) and (science or biology or chemistry or physics or earth science)
7 Limit 6 to (LA:PY = ENGLISH) and ((PT:PY = CASE-STUDY) or (PT:PY = CLINICAL-TRIAL) or (PT:PY = COLLECTED-WORKS) or (PT:PY = CONFERENCE-PROCEEDINGS-SYMPOSIA) or (PT:PY = EMPIRICAL-STUDY) or (PT:PY = EXPERIMENTAL-REPLICATION) or (PT:PY = FOLLOWUP-STUDY) or (PT:PY = INTERVIEW) or (PT:PY = JOURNAL-ABSTRACT) or (PT:PY = LITERATURE-REVIEW-RESEARCH-REVIEW) or (PT:PY = LONGITUDINAL-STUDY) or (PT:PY = META-ANALYSIS) or (PT:PY = PROGRAM-EVALUATION) or (PT:PY = PROSPECTIVE-STUDY) or (PT:PY = RETROSPECTIVE-STUDY) or (PT:PY = TREATMENT-OUTCOME-STUDY)) and (PY:PY = 1980-2002)

2.2.4 Social Science Citation Index (SSCI)

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

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

Appendix 2.3: Journals handsearched

The following key journals were handsearched for potentially relevant papers:

Journal of Biological Education
Journal of Chemical Education
Research in Science and Technological Education
Research in Science Education
Studies in Science Education

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

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

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

A2. Status
Published
In press
Unpublished

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

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

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

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

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

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

A8. Programme name (Please specify.)

....

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

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

A11. Sex of learners
Female only
Male only
Mixed sex

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

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

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

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: Enhanced data-extraction tool

Guidelines for additional questions to be added to EPPI-Centre data extraction tool (EPPI-Reviewer)

E1-E6	**Normally not applicable**	
E7	**Study design summary**	
E7	1	*What is the rationale for the study design (e.g. reasons for different components/stages of research; purposes of different methods or data sources; justification of time frame; study location; etc.) in relation to the study aims? And what limitations of the study design are reported?*
E7	2	*What is the status and role of the researcher in relation to the study?*
E7	3	*What are the key characteristics of the sample (sample size/case numbers for age, gender, ability, socio-economic background, any other features of note)?*
E7	4	*How does the sample selected relate to the population of interest (e.g. typical, extreme, diverse constituencies, etc.)? What rationale is given for the selection of the target sample (basis for in/exclusion, discussion of sample size/case numbers, selection of settings, etc.)*
E7	5	*What information is given about the strengths and weaknesses of data sources and methods?*
F1-F4	**Normally not applicable**	
I5-I7	**Normally not applicable**	
I9	**Important features of data collection**	
I9	1	*What information is given about data-collection methods (e.g. for interviews/conversations: audio or video recordings; for field notes: conventions for recording events and/or distinguishing between description and comment/analysis)?*
I9	2	*What information is given about data-collection instruments (e.g. interview schedules, observation sheets, field notes routines, diary keeping instructions, etc.)*
I9	3	*What information is given about measures for increasing trustworthiness of the data collected (e.g. influence of fieldwork methods/settings on the nature of the data collected; features of data indicating depth, detail, richness; etc.)*
J2-J7	**Normally not applicable**	
J8	**Important features of analysis**	
J8	1	*What information is given about how descriptive analytic categories, classes, typologies, etc. have been generated and used (may be descriptive or constructed categories)?*
J8	2	*What information indicates the portrayal of the context of data sources (e.g. reports take cognisance of historical developments and social characteristics of study sites and settings; participants' perspectives are placed in personal contexts; etc.)?*
J8	3	*What information indicates diversity in the data (e.g. multiple perspectives, alternate positions, negative cases, outliers, exceptions)?*

J8	4	*What information is given about patterns of association or linkages with divergent positions or groups in the data, including possible explanations?*
J8	5	*What information is given to indicate measures for increasing trustworthiness of the analysis process (e.g. exploration of the detail, depth and complexity; exploration of contributors' terms, concepts and meanings; discussion of explicit and implicit explanations, discussion of underlying factors/influences; discussion of patterns of association/conceptual linkages within data; presentation of illuminating extracts/observations)?*
J8	6	*What information is given about any corroborating evidence used (e.g. other data sources or research evidence used to support or refine findings)?*
J8	7	*What information is given about the generation of criteria for effectiveness or impact? What information is given about how evaluative judgements have been reached?*
J8	8	*What reflection is included about unintended consequences of the intervention, and about the implications of changes in the research design?*
J9	**Normally not applicable**	
M1	**Ethical consideration**	
M1	1	*What information is provided to indicate sensitivity to research contexts and participants when presenting the study to participants (consent procedures; information provided on anonymity of participants/sources and confidentiality of data; information to be supplied at the end of the study; potential problems with participation)?*
M1	2	*What are the indications of sensitivity to research contexts and participants in the reporting of the study (anonymity of informers, etc.)?*
M5	**Normally not applicable**	
M7	**Normally not applicable**	
M10	**Respond in terms of relatability** *(the extent to which the reader can judge the communalities between the research context and their own/other situations)*	

APPENDIX 2.6: Weight of evidence indicators

Review question: What is the nature of small-group discussions aimed at improving students' understanding of evidence in science?

Weight of evidence B: appropriateness of research design and analysis for addressing the question *of this specific systematic review*			**Weight of evidence C:** relevance of particular focus of the study (incl. conceptual focus, context, sample and measures) for addressing the question *of this specific systematic review*			**Weight of evidence D:** Taking into account M11, B and C: what is the overall weight of evidence this study provides to answer *this review question*?
high (3)	medium (2)	low (1)	high (3)	medium (2)	low (1)	If equal weighting of M11, B and C, each weighted across the range as low (1), medium low (2), medium (3), medium high (4) and high (5)
For the RQs relevant to the review			**For the RQs relevant to the review**			
sampling frame great detail of nature of sampling frame **actual sample** comparison between/ within group in design **context of SGD** greatly detailed description **data collection** high trustworthiness of data collection methods **data analysis** high trustworthiness of data analysis methods	reasonable detail of nature of sampling frame comparison between/ within group in findings only reasonably detailed description medium trustworthiness of data collection methods medium trustworthiness of data analysis methods	little detail of nature of sampling frame no comparison hardly any detail of context provided low trustworthiness of data collection methods low trustworthiness of data analysis methods	**focus of intervention** understanding evidence in science is main focus of intervention **focus of study** nature of discussion is explicit independent variable **measures** highly appropriate for testing nature of discussions directly **breadth** reports broad range of nature of discussion **situation** highly representative of SGD in classrooms	understanding evidence in science is one of several foci of intervention nature of discussion is a major discrete element of study mildly appropriate for testing nature of discussions directly reports narrow range of nature of discussion less representative of SGD in classrooms	understanding evidence in science is tangential to intervention nature of discussion is tangential interest of study appropriate for testing nature of discussions indirectly reports nature of discussions only indirectly not representative of SGD in classrooms	Sum total and classification for D: 3–4: low 5–7: medium low 8–10: medium 11–13: medium high 14–15: high

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

Appendix 3.1: Types of study included in the systematic map

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

Table A lists the 14 reports of descriptive studies.

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

Tables C and D list the reports of the 23 naturally-occurring and 25 researcher-manipulated 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 = 14)

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

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

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

		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
1777	Woodruff and Meyer, 1997	✓	✓		✓	✓				
1678	Vellom *et al.*, 1995	✓	✓		✓	✓			✓	
1389	Roth and Roychoudhury, 1992	✓	✓		✓	✓			✓	
693	Hogan, 1999a	✓	✓		✓	✓		✓	✓	
1632	Tsai, 1999	✓		✓		✓		✓	✓	
1824	Osborne *et al.*, 2002	✓	✓	✓	✓	✓	✓	✓		✓
1457	Seiler *et al.*, 2001	✓	✓	✓	✓			✓	✓	
1514	Solomon, 1992	✓	✓	✓	✓			✓	✓	✓
2049	Russell *et al.*, 2003		✓		✓	✓				
1544	Stein, 1997		✓			✓				

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

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

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

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

		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
1218	Pederson, 1992	✓	✓	✓	✓			✓		
692	Hogan, 1999b	✓	✓	✓	✓	✓		✓		
1471	Sherman and Klein, 1995a		✓	✓		✓		✓	✓	
258	Chang and Lederman, 1994		✓		✓				✓	

Appendix 4.1: Details 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) were 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 whole actual dialogues of students working on their tasks

	are presented. • 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 use of four main categories of Dialogue Turns • 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 all data collected from student dialogues
Summary of results	• Topic domain (sound): episodes of epistemic dialogue were closely related to levels of description, different conceptual perspectives and double meanings in the domain, and 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 use of volunteers reduces generalisibility. Reporting the effect of the different features of CONNECT for the different dyads individually, rather than across the dyads, would strengthen the findings

Weight of evidence B (appropriateness of research design and analysis)	**Medium** Some detail of how groups were formed. Little detail about the sample (school/class). Comparison between groups (only two) in findings. Much detail of task/prompt, less so on group members. High trustworthiness for data collection (pilot pair discussions and data logs with verbatim exchanges), and data analysis (development of detailed coding scheme with illustrations and quotes from data; reliability checks across three researchers).
Weight of evidence C (relevance of focus of study to review)	**Medium-high** Main focus of intervention is dealing with/understanding of evidence. The focus of the study is on the nature of epistemic dialogue AND the characteristics of the software. Verbatim records of discussions are highly appropriate measures. Good breadth of nature of discussions (argumentation and explanation). Situation unrepresentative of classes (volunteers, after class, group members in two 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. Exam taken co-operatively at the end of the first phase allowed students to demonstrate their ability to use two basic genetics models. Genetics data conflicting with these models are provided and generated. Group presentations of revised models are presented and critiqued. 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 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 because this was an elective course 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** Great detail of sample frame and sample (school/teacher characteristics and nature of class selected). Some comparison of observed strategies of different groups but not linked to group characteristics. Great detail of intervention task/stimulus but not of group members. High trustworthiness of data collection (triangulation, pilot testing). Detailed construction of analysis scheme but categories, and narrative descriptors not presented. Few quotations.

Weight of evidence C (relevance of focus of study to review)	**Medium** The understanding of evidence (model revision) is the major focus of the intervention. The nature of the group discussions is a vehicle for the main interest of the study, i.e. the different types of knowledge activated for model revision. The measures (classification schemes for types of knowledge) are incomplete and are hardly appropriate for the nature of discussions. It reports the nature of discussions only indirectly. Takes place in an elective class, so some limits to representativeness.
Weight of evidence D (overall weight of evidence)	**Medium**

Hogan K (1999a) Sociocognitive roles in science group discourse. *International Journal of Science Education* 21: 855–82.	
Country of study	USA
Details of researchers	Not stated
Name of programme	Not applicable
Age of learners	Eighth grade USA (aged 13 to 14)
Type of study	Exploration of relationships
Aims of study	To explore factors which limits and promotes students' learning in small-group discussions
Summary of study design, including details of sample	• A long timeframe is justified for verifying the constancy of roles. Interviews (probing personal perspectives of science learning) are sensibly used to compare with observed roles, and groups' dynamics. Observations of eight groups within same class is justified by the need to compare dynamics and roles for constant task. • The researcher is a non-participant observer, keeping field notes. • Discourse in eight groups of three students each from four classes is used as evidence for individual's roles and group dynamics. Each group is heterogeneous for ability, and uses some friendship ties. Gender and race of students provided, but no socioeconomic background. Twelve of these students were interviewed individually. • No indication of how eight groups are selected, or how they relate to the population in the four classes. • No discussion of strength and weaknesses of data sources or methods. • Details of sample: 24 students in eight groups of three. Group is unit of analysis.

Methods used to collect data	• One-to-one interview • Observation: audio- and videotaping, with field notes of class observations • School/college records • Other documentation: field notes
Data-collection instruments, including details of checks on reliability and validity	• Observation by researcher of three to four sessions per week over 12 weeks for four classes. No details on format of field notes. • 9–10 interactive sessions per group as planned by the teacher, totalling 73 sessions. • Audio- and videotaping using normal conventions: verbatim transcription and narrative description (for off-task dialogue). • Key questions for semi-structured interview schedules are provided. The role of the use of prompts is justified. Stimulated recall methods used in second interview. Guidelines for field notes are not provided, but these are used only occasionally for triangulation. • The method of group composition (not fully friendship based) influences the data, and this is highlighted by the author. The fact that two classes of two teachers were used would allow for identifying striking teacher influences (not reported).
Methods used to analyse data, including details of checks on reliability and validity	• Ethnographic interaction analysis was used on large sections of discourse for identifying and interpreting patterns of group interactions and roles played by individual students. Results of task used in interviews were used as the basis for follow-up questions to gain perspectives that might underlie learning preferences. Later interviews gained data on memorable learning events partly prompted by video (stimulated recall). Descriptive profiles of students' perspectives of learning science were constructed from the entire interview transcript. • Rich personal contexts: display (including through quotations) of interpersonal relationships • Data presented in a rather convergent manner. No attempt to disprove assertions. • Study is about identifying associations between group interactions, individual roles and their perceptions of learning science. • Increased trustworthiness: triangulation (audio data and field notes) has been applied and presented; meanings of individuals maintained and interpretation presented; extensive group exchanges used for illustrating deep and surface reasoning. • Some reference to outcomes of earlier study of same author, and in the discussion section refers to other literature. • There is no intention to evaluate the intervention.
Summary of results	• Students took eight sociocognitive roles in group reasoning processes: four positive (promoter of reflection, contributor of content knowledge, creative model builder, mediator of social interaction and ideas) and four negative roles (promoter of distraction, of acrimony, of simple task completion, reticent participant) • These roles are persistent over time. • The roles of leader and helper in investigative group work (identified in other literature) do not emerge in a mental-model building context. • Deep (as opposed to surface) reasoning in a group is related to the presence of at least one promoter of reflection. • Personal preferences of science learning and personality traits influence role adoption. • Students with different personal attributes may take identical roles depending on the group context.

Conclusions	• Assigning managerial or prosocial roles (e.g. reflector, regulator, questioner, explainer) as in collaborative learning theory may be counter productive for ill-structured intellectual tasks. Intellectual roles should then be allowed to emerge naturally. • Guiding students towards taking constructive roles may be achieved through metacognitive training – that is, knowledge about the nature of collaborative learning, effective group learning strategies, and awareness of what constitutes progress. • If the intention is to develop meaning-making abilities and higher-order thinking in small groups, friendship groups are more effective. If the intention is to improve the ability to work with peers, this may be better served through shorter and less intellectually demanding tasks.
Weight of evidence A (trustworthiness in relation to study questions)	**Medium** The sample seems well detailed, but the sampling method for the groups is unclear. Data-collection methods are very appropriate for the research questions. Data analyses of the transcripts for both the student roles and the group activities are trustworthy. The steps for the analysis of the interviews are less convincing, and seem to start from a comparison with the individual's role predetermining the student profiles for their perceptions of science learning. The findings are very difficult to spot.
Weight of evidence B (appropriateness of research design and analysis)	**Medium-high** Some detail of the sample frame (school/class/teachers), but lacks justification for selection of eight discussion groups. Good comparison between eight groups on the basis of pre-determined surface/deep reasoning ability, and of findings on individuals' roles. Rich description of context of SGD (tasks/strategies and personal conflicts, gender, reasoning styles). Data collection highly trustworthy with multiple sources, detailed interview strategies. Data analysis uses grounded theory approach. Quotations extensive. No indication of independent coding.
Weight of evidence C (relevance of focus of study to review)	**Medium-high** Focus of the intervention is sense-making model construction, so directly on dealing with evidence. Role of participants, not the nature of discussion is the central interest in the study. Measures lead to classification in deep/surface reasoning, although this is only part of the study. Breadth of nature of discussion considerable. Natural classroom situation.
Weight of evidence D (overall weight of evidence)	**Medium-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
Aims of study	To evaluate the effect of an intervention stressing the metacognitive and group strategic aspects of co-constructed knowledge 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: • 2 x 2 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 metacognitive knowledge about collaborative reasoning and ability to articulate their collaborative reasoning processes compared with students in control classes. • This enhanced metacognitive 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 by 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	Medium Not any higher because of general lack of information on the rigour of the qualitative component of the study.

study questions)	
Weight of evidence B (appropriateness of research design and analysis)	**Medium** Sample frame well described (school/teacher/class, even some group profiles). Comparison of groups on the basis of initial differences in treatment. Hardly any detail of context of discussions (task/group composition and characteristics). Data collection good but on samples of discussion, there is no clarity how selected. High trustworthiness of data analysis based on existing categories, but short on illustrative quotes, and without indication of double-coding.
Weight of evidence C (relevance of focus of study to review)	**Medium** The intervention focused on collaborative reasoning (thus understanding of evidence). The main focus of the study is on the effectiveness of the intervention; the nature of discussions is a secondary focus. Measures of nature of discussions is only reported indirectly (deep/surface). Very limited breadth of nature of discussion. Classroom setting.
Weight of evidence D (overall weight of evidence)	**Medium**

Jiménez-Aleixandre MP, Pereiro-Muñoz C (2002) Knowledge producers or knowledge consumers? Argumentation and decision making about environmental management. *International Journal of Science Education* 24: 1171–1190.	
Country of study	Spain (Galicia)
Details of researchers	Not stated
Name of programme (if applicable)	Part of the Reasoning, Discussion and Argumentation (RODA) project
Age of learners	Evening shift, so larger age range (17 to 21) than equivalent school population (15 to 16)
Type of study	Exploration of relationships: The relationship is between features of argumentation and decision-making process on the one side, and novice (students) or expert status on the other.
Aims of study	From abstract: The purpose was to study the components of knowledge and skills needed to reach a decision in socio-scientific contexts, and to identify them in classroom discourse. Page 1173 describes the purpose as being to explore whether the students in a problem-solving context can act as knowledge producers reaching decisions about environmental management, and to compare their steps and arguments to the ones of an external expert.
Summary of study design, including details of sample	• Six groups of four to six members (changing composition by jigsawing twice) are presented with an environmental problem and two possible solutions, one of which is discussed in the expert's project report. Decision-making is based on expert project descriptions, field trip information, debate between experts and students, and group study of different science aspects of the problem. Each group produced a written report, assessing the proposed project and its predicted impact. These reports were debated in the whole class. • Environmental issue is justified as authentic for these students. The staging (accessing different sources of information) is directly related to the RQs. No limitations of design reported. • One of the two researchers is the teacher. There is no discussion of the implications of this double role. • Not very representative sample: class of 38 students in evening shift with the age range 17 to 21. Authors report that students are either working during the day, or wishing to take the course at a slower pace. No information about gender, ability, socio-economic background of members of each discussion group. No information how school or class was selected. • Not very representative sample: large group, in evening shift (probably unstable in composition) with wide age range. No rationale provided for the selection of this school or class – most probably opportunistic as researcher was teacher. • No detail of strengths/weaknesses of sources and methods. • Details of sample: 38 students and two experts
Methods used to collect data	• One-to-one interview with the expert • Observation: audio-/video-recordings and field notes from observations

	• Self-completion report or diary: field notes from an external observer; students' individual and collective reports and other work collected in their portfolios
Data-collection instruments, including details of checks on reliability and validity	• Data were collected through audio- and video-recordings of small group discussions (all six groups?) and plenary reporting sessions. No information about field note report format, or student report format. • No formal data-collection instruments were used. • Data collected from multiple sources could provide the opportunity of triangulation, but this is not mentioned. There is some critical reflection on the fact that two solutions were provided, predetermining a focus on the choice between these two, rather than creating an alternative solution.
Methods used to analyse data, including details of checks on reliability and validity	• Two tools used for analysis of verbatim transcripts of student discourse. Toulmin's discourse schemes are used to identify warrants, but the classification of warrants emerges from the data; this provides the basis for the comparison of student and expert warrants. Walton's five questions about expert argument were collapsed into two questions. Discourse was analysed to identify these two issues. The classification of generated criteria emerged from the data. Purposely the views of the engineer (not the ecologist) were compared with those of the students, since most student groups opposed the project solution (proposed by engineer) and would use similar arguments as the ecologist. • The context of the students/expert are presented faithfully only to the extent that many extensive quotations are provided. • There is no specific attempt to identify negative cases, or identify patterns of association. • Extensive transcript illustrating the ambiguity of the use of warrants or backings (p 1177). In general, the rich quotations provide good flavour of the discussions. • References to research are mainly methodological, not corroborating the evidence. • There is no mention in analysis of field notes or written student reports. • Statistical treatment is limited to frequency counts.
Summary of results	• Students constituted a knowledge production community by combining ecological concepts with technical information. • In reaching conclusions, students applied conceptual knowledge at more than a surface level. • Warrants used by students and experts showed concordance indicating that students were not just passive consumers of knowledge. • While warrants used by students and experts covered the same concepts, they were used to support different claims; expert claims were often supported by multiple warrants, and students' by single warrants. • Initial criteria used by students to evaluate solutions to the environmental problem were general but, as they progressed with the study, they became more refined and specific. • Initially students did not consider themselves experts enough to weigh arguments. Only the identification of conflict between evidence and (other) expert opinion allowed some student groups to assign expertise to themselves. • Student decisions were not based on conceptual understanding or scientific evidence alone, but value judgements played an important role. • The weight of value judgements in decision-making depended on students' ability to reason in opposite domains (ecology-economy); students ranked ecological and political values high; the expert ranked economic and practical values.

Conclusions	See findings
Weight of evidence A (trustworthiness in relation to study questions)	**Medium-low** Data collection through classroom interactions is appropriate, but there is no feel of the coverage of these data. The possibility for triangulation (from student written work) has been missed. The research questions have no longitudinal component, but some of the findings are reported as such. The task is authentic, and admirably complex. The analysis methods for views of authority of information sources and the development of criteria for deciding solutions are not convincing. The sample (evening class) is untypical.
Weight of evidence B (appropriateness of research design and analysis)	**Medium-low** Little detail of the school/class (sample frame). Hardly any comparison (other than mentioning these) for differences in groups' conclusions or arguments, apart from 'authoritative' group. Some detail of group discussions. Data-collection methods are appropriate, but lack triangulation. Two of the three data analysis frameworks are unclear.
Weight of evidence C (relevance of focus of study to review)	**Medium-high** Understanding of evidence (argument, authority of evidence sources, criteria for judgements) is the prime focus. Discussions within groups one of the foci (apart from comparison with expert argument). Measures (audio/video recorded discussions, observations) appropriate. Good range of the nature of discussion. Untypical situation (evening class, large age range, some adult learners).
Weight of evidence D (overall weight of evidence)	**Medium**

Jiménez-Aleixandre M, Rodriguez AB, Duschl RA (2000a) "Doing the lesson" or "doing science": Argument in high school genetics. *Science and Education* 84: 757–792	
Country of study	Spain (Galicia)
Details of researchers	Not given
Name of programme	Standard lessons on Mendelian genetics
Age of learners	Grade 9 (aged 14 to 15)
Type of study	Exploration of relationships
Aim of study	To explore the conversational dynamics in the form of argumentation patterns and epistemic operations students perform when solving a problem in science classes (p 759)

Summary of study design, including details of sample	• The discourse patterns in SDG in more frontal class (occasionally in small groups) is contrasted with actual small-groups discussions centred on a problem-solving task: why do farm chickens have yellow feathers whilst wild ones are spotted? The two last lessons with two progressively focussing tasks culminate in whole class group presentations and discussions. • The stages are prescribed by the learning sequence. The researchers purposely chose a class midway between student- and teacher-centred – used to SGD work and freely contributing to whole class discourse, with teacher willing to use learner-centred methods, but lacking some of these skills. Relevant teaching/learning strategies are illustrated by the data. • Researchers were university academics from the University of Santiago de Compostela and King's College, London. They proposed the intervention task. • One whole class of ninth graders in a Spanish public high school (six groups of four students) performed SDG problem-solving tasks. The discourse of one group of four girls is the focus of the study, together with the whole class discussion of all six groups. Comprehensive school in a medium-sized town in North Western Spain. Little information on characteristics of students (ability, socio-economic background, language fluency). The biology teacher had five years teaching experience. • No rationale is given for selection of sample. From the whole-class discussion, it seems that the focus group was reasonably representative of the whole class. • No discussion of strengths/weaknesses of data sources/methods • Details of sample: one class; number of students is not stated.
Methods used to collect data	• Observation combined with audiotaping of group discussions.
Data-collection instruments, including details of checks on reliability and validity	• Audiotaping is self-explanatory. No description of observation protocol, or routine of field notes taken as a result of the observation. • Audio-recordings of group discussions. In small groups, individuals were identifiable, but in the whole class, only groups were identified. No information about observation methods. • No information about data-collection instruments, other than the task table students needed to complete. Verbatim transcriptions of recorded group discussions were made. • Verbal group interactions were captured in their entirety, providing good depth and detail. Observations provided opportunity for triangulation and adding non-verbal communication, but this is not reported.
Methods used to analyse data, including details of checks on reliability and validity	• Two classification schemes for classroom discourse have been taken from the literature. No reasons were given for the choice of analysis frameworks. The third scheme (epistemic operations) was developed, based on several theoretical classifications and refined on the basis of the data. • Audiotapes were transcribed and sentences broken into 'units of analysis' (no definition given of this term). Units were examined for evidence of 'doing the lesson' or 'doing science'. Units were coded with reference to work done by Bloome *et al*. Units were examined for evidence of 'talking science'. Units were coded as 'argumentative operations' or 'epistemic operations' and with reference to Toulmin's work in the former case and other work in the latter case. • Contextualised analysis is provided through details of gender and ability of students in the small group. Some tensions between group members come through. Some presentation of statements, interpreted differently by different participants. However, for the whole class discussion, the data were not interpreted in their context, apart from portraying the development of 'camps'. • The analysis framework appears to have been developed, and examples selected from the data to illustrate aspects of the framework.

	This is not really a paper where theory has emerged from the data. So the notion of diversity in the data only applies to variety within the examples given, where some extracts illustrate aspects of the framework better than others. • Patterns of association: N/A for pre-determined analysis framework • Trustworthiness of the analysis is high: Several examples are provided where multi-interpretation is possible, thus increasing the trustworthiness. Authentic terms used by students are reported and interpreted. Data were compared across the observation period, and across SDG and whole class interactions. However, the choice of extracts to present has been made by the authors, who are likely to have picked examples which best illustrate aspects of the analysis framework. Also, at least one of the authors is very interested in promoting the teaching of argumentation skills in school science. This does not make the analysis untrustworthy, but might influence choice of data to present. • In the discussion, supporting and conflicting evidence from other studies are used to put data analysis in perspective. • No evaluative judgements are intended, as there have been no attempts to teach pupils argumentation skills.
Summary of results	• A large proportion of discourse statements relate to ‘doing the lesson’ (interaction referring to the rules of the task, or to perceived features of science classrooms). Later in the discussion, statements indicating ‘doing science’ increase. • 35% of arguments in SGD are claims, 20% warrants, 10% call on data and 5% are backings. Rebuttals and qualifiers only occurred in the plenary discussion. • Arguments are developed by a subset within the group. Although agreement is offered (for social reasons), deviating personal opinions may still persist. • In line with the causal nature of the task, most epistemic operations reflected causality. In addition, there were differences in the nature of the ideas developed by students, with some evidence of ‘anthropocentrism’ – that is, students not appreciating that theories of heredity needed to be able to be applied consistently to a range of organisms, rather than have different theories for different organisms.
Conclusions	• If the ability to develop arguments is set as a learning goal in science, this will not happen during normal instruction; specific inquiry-focused tasks need to be provided. The process of inquiry specifically is relevant for developing the epistemic goals (i.e. understanding the structure of knowledge). • Given a classroom environment conducive to discussions and providing and defending opinions, even untutored students will use a number of operations (argumentative and epistemic). A variety of tasks and approaches will need to be provided for students to solve problems, discuss scientific issues, relate data and offer explanations (a bit circular: they used variety successfully, therefore variety is needed). • Hypothetical, unquestionable data (provided by the teacher) will generate different patterns of argument compared to empirical, uncertain data.
Weight of evidence A (trustworthiness in relation to study questions)	**Medium-high** The sample selection of teacher and class is fully justified and appropriate for the study. The reasons for selecting the (one) small group is not clarified. Data-collection methods (audiotaping) are highly appropriate, but a report of the use of observational data would increase the trustworthiness (triangulation and increased context thickness). The analysis schemes have been appropriately applied from previous studies or developed from the data. The extensive quotations and detailed inferences provide good context to the presentation

	of results, but feel as if they are selected to illustrate the framework positively. The report of the traditional classroom interactions provide high relatability to the findings.
Weight of evidence B (appropriateness of research design and analysis)	**Medium-high** Reasonable detail is provided of the school and class. Minimal comparison within group or between groups. Highly detailed description of SGD. Audiotaping has high trustworthiness for data collection (could have been even higher if complemented by observation data). Analysis methods are based on two existing classifications and one well-developed grounded classification.
Weight of evidence C (relevance of focus of study to review)	**High** Main focus is the understanding of evidence. The nature of the discussion is explicit variable. The measures are very appropriate for testing the nature of discussions. Reports on broad range of aspects of discussion. The classroom situation is natural.
Weight of evidence D (overall weight of evidence)	**Medium-high**

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**	
Country of study	USA
Details of researchers	Teacher/researcher (for PhD)
Name of programme	No name of the programme is mentioned, but the intervention is the same as that used in the study by Finkel (1996). Both make use of Jungck and Calley's software (Genetic Construction Kit) to generate genetic populations and crosses.
Age of learners	17 to 18
Type of study	Exploration of relationships: The relationship between 'success' (defined in its particular way) and problem-solving strategies.
Aims of study	The aim of the study is to describe the strategies students use in model-revising problem-solving in genetics, with a view to informing the design of model-based instruction. It looks like the focus involves presenting students with 'discrepant events' (i.e. things which do not fit) and requiring them to modify their ideas.
Summary of study design, including details of sample	• The problem-solving strategies used during three sessions by two student groups (one successful the other unsuccessful) are compared. Being successful means, in this study, that groups are satisfied with the revised models they have come up with (success is not based on the degree to which the new models are scientifically correct). No justification for selecting these students/this school. No

	limitations are reported. • The researchers are each 'interested parties' (one is participant observer and teacher of both groups; the other is curriculum designer of the intervention). Not commented upon. • Out of four focus groups, two groups of three students each (three males; two males and one female) were selected. The two groups were selected retrospectively for their diversity in being successful in revising given genetics models to take account of anomalous genetic data. No indication of students' ability or cultural, racial, language or socio-economic background. • School draws from a combination of suburban and rural areas, some (not specific) students were interested in pursuing study of science (p 465). Authors claim groups studied were representative of successful and unsuccessful groups, but no detail is provided of what this means, except what is implied in the reporting of the data. • No information or discussion of strengths and weaknesses of data sources, but an extensive discussion of limitations of the method: the authors recognise that differences in problem-solving strategies may not be the only factors contributing to differences in 'success'. They list three probable alternative factors.
Methods used to collect data	• Self-completion report or diary • Audiotapes of discussions
Data-collection instruments, including details of checks on reliability and validity	• Data collection through continuous audiotaping of group discussions and whole class group presentations and through student written work including lab books (descriptions of problems being solved, experiments performed, protocols used, models considered, final model produced) and course journals (responses to questions posed by teacher). The audiotapes provided the main data source. • No actual data-collection instruments were used. • Although the audiotapes provided the main data source, triangulation with other sources was probable, thus increasing the trustworthiness. Little is written about the characteristics of the group members, nor about the group dynamics.
Methods used to analyse data, including details of checks on reliability and validity	• Discussion transcripts analysed for (a) points where anomalies were recognised, and (b) strategies used for model revising (p 468). • The coding used and extended categories from previous studies by one of the authors. Examples of dialogue provided for 'anomaly resolution' and 'model assessment' (p 469, Table 4). • No information is given about the content of the sources. • One successful and one less successful group provides some diversity in the data. • In the context of the focus of the discussions, different student perspectives are reported, organised around repeating or contrasting patterns. Little in the way of explanations. • No details were given of the analysis of data from student records and products; therefore no depth was added to the analysis of the transcript data. • No corroborating evidence is provided, even at the level of relating findings to work referred to in section on 'theoretical background'. • If criteria for impact mean that students had grasped the principles of inheritance, these were implicit in the nature of the students' discussion. • No mention is made of unintended outcomes, and no reflections are made on implications for research design.
Summary of results	Many of the findings are very specific to the science being studied, and remain virtually at the level of presenting the data.

	General findings are: • Groups were able to recognise anomalies in models, but the less successful group did not always do so, and sometimes chose to ignore them. • The more successful group was more methodical in testing its revised models. Successful model revision is helpful if: • the rate of anomaly detection is increased by comparing with both given models • identified anomalies are all used for developing tentative models • proposals and assessment of new models take account of both given models (not only one) • the tentative models are assessed on their power to explain and predict the cross results at phenotypical and genotypical levels • strategies are used systematically, such as crossing like phenotypes, using Punnett squares, and constructing generations of organisms
Conclusions	• Discussion helps students articulate revised models. • The authors plan to build in some specific instruction on problem-solving into the next phase of the work.
Weight of evidence A (trustworthiness in relation to study questions)	**Medium-low** Purpose of the research is unclear: Comparisons are made for strategies of students who are satisfied or dissatisfied with the revised models they have constructed. The study draws on a narrow range of genetics research. The data-collection methods seem reasonable, although video-recording groups would have been helpful in focussing student talk in such interactions. Data analysis seems fine, but the results do not get much further that presenting the data. We know too little about the students, teacher, class and school to make the findings relatable. A limitation is the very small sample size.
Weight of evidence B (appropriateness of research design and analysis)	**Medium** No detail of sampling frame. Comparison between two SGD as part of the design. No detail of context of SGD. Data collection reasonably trustworthy. Data analysis reasonably trustworthy.
Weight of evidence C (relevance of focus of study to review)	**Medium-low** Understanding of evidence is the main focus of intervention. Nature of discussion tangential to study. Transcription of group discussions appropriate data source for SGD, but measures not appropriate for testing nature of discussions, but of understanding evidence. No breadth of measures for testing the nature of SGD. Computer lab is reasonably representative for class situation.
Weight of evidence D (overall weight of evidence)	**Medium-low**

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
Aim of study	To investigate the use of reasoning strategies through a collaborative writing task in order to generate meaningful scientific models and the evidence for improvement in students' reasoning discourse
Summary of study design, including details of sample	Pre- and post-intervention clinical interviews with four individual students regarding conceptual knowledge Two single-sex pairs underwent the intervention and generated collaboratively a report for two laboratory activities. The domain-specific knowledge for one activity was low, for the other high. Reasoning strategies in interactions between pairs were video-recorded, and in individual and joint written products collected. The types of reasoning strategies resulting in conceptual change were identified. For paper 2, no interviews were used, and three pairs were involved. The types of reasoning strategies used were classified and their development over a three-month period traced. Actual sample: Paper 1: two pairs, four students. Paper 2: three pairs, six students.
Methods used to collect data	• One-to-one interview: Pre- and post-intervention clinical interviews • Observation: Video-recorded pair interactions (two cameras!) • Self-completion questionnaire: Written collaborative report of laboratory activity. Written individual prior knowledge and predictions. • School/college records • Other documentation: Researcher's field notes
Data-collection instruments, including details of checks on reliability and validity	• Sample of a reporting guideline is appended to paper 2. • No interview schedule is provided, but relevant interview responses are reported verbatim. • Checks on reliability: Triangulation of data sources (field notes, video footage, written records) increases reliability. • Checks on validity: This is an interpretive study, so the emphasis is on contextual validity: extensive details are provided of the type of characteristics of students and the process of their involvement, the teaching procedures, and the context of the specific task being

	focused on. Some more detail on the general environment in the school would have been useful. • One task was used for development of a pilot collaborative report.
Methods used to analyse data, including details of checks on reliability and validity	• This is an interpretive study. Descriptive analysis: The domain-specific understanding in pre- and post-intervention interviews has been described according to the nature of concepts – accepted major types of misconceptions are used as classification. A constant comparative method was 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* • RQ 1: 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. co-ordinating all mutually consistent knowledge propositions into a coherent model. • RQ 2: 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* • RQ 3: 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). • RQ 4: The greatest improvement in reasoning discourse occurs in pairs which 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 medium-high trustworthiness.
Weight of evidence B (appropriateness of research design and analysis)	**Medium-high** Limited detail of the sample frame, but some of the discussion group members and the selection method. Comparison between groups (difference in conceptual understanding) in design. Detailed context of SGD (tasks, strategies, intra-group tensions). Good contextual validity of data collection. Good data analysis (triangulation, modification of Kuhn's framework).

Weight of evidence C (relevance of focus of study to review)	**Medium-high** Focus of the intervention is scientific reasoning skills rather than on dealing with evidence. Nature of discussion is the central interest in the study. The measures (dialogue contributions) are fine but presentation has limited detail. Reports on four major aspects of the nature of discussion. Naturalistic classroom situation.
Weight of evidence D (overall weight of evidence)	**Medium-high**

Kurth LA, Anderson CW, Palincsar AS (2002) The case of Carla: dilemmas of helping all students to understand science. *Science Education* 83: 287–313	
Country of study	USA
Details of researchers	Not given
Name of programme	Modified module on density from 'Colored Solutions' curriculum
Age of learners	11 to 12
Type of study	Exploration of relationships
Aim of study	To create a classroom community that functioned as a discourse community, sharing some, but not all, of the characteristics of adult scientific communities
Summary of study design, including details of sample	No stages of research are specified; the timeframe had to fit with the teaching schedule. Study rationale: a desire to apply an interpretive framework (see p 291) to analyse group discourse. (Note: the framework looks quite interesting.) The nature of discourse analysis dictates that there will be small sample sizes. No mention at all is given to any possible limitations in design. The role of researchers is not stated. Implicit is that the researchers gathered the interview and observation data. The focus is on four pupils from an intact class of 29: two girls, two boys, two European Americans, one Mexican American and one African American Groups were constructed by the teacher to 'maximise diversity' for gender, ability, race and cultural background. The school drew on a wide range of social class and ethnic backgrounds. No information is given about the strengths and weaknesses of data sources and methods.

Methods used to collect data	• Curriculum-based assessment • Observation • Self-completion questionnaire • Self-completion report or diary
Data-collection instruments, including details of checks on reliability and validity	• Use of two simultaneous fixed video cameras for class interactions. Single fixed camera for group discussions, backed up by microphone on bench. • No details are given about data-collection instruments (individual interview protocols, or tests). • No detail on context of data collection – but task focus on making, observing, explaining stacks of colour solutions. • No details are given about data-collection methods. • Gathering both interview and observation data could be described as enhancing the trustworthiness of the data (triangulation). • The way in which the data are presented in the paper suggests a high level of detail. • General comment: the data collection methods are appropriate to the type of study being undertaken, where the emphasis is on gaining a rich, in-depth picture of a situation. However, the weakness is that the methods are not described or justified in any detail.
Methods used to analyse data, including details of checks on reliability and validity	• The analysis was done with reference to an interpretive framework of four concepts central to discourse: polyseny, defining the floor, privileging and intersubjectivity (see pp 291–293). • Participants are commented on consistently from their personal perspectives. • Presentation of group discourse with reflective interviews provides multiple perspectives. No attempt to search for negative cases, but the group is rather small for this. • Patterns of association for positions of different positions are difficult to spot in such a small group. • Reasonably trustworthy: There is a great detail of discourse; with appropriate comment, often from the perspective of several participants; contributors' terms are explored; implicit/explicit meaning discussed. • There are points in the presentation of the descriptive data (e.g. p 295, p 297) where the findings are related to other studies. Generally, this happens when the study findings resonate with those of other studies, rather than when they differ. • No criteria for judging effectiveness have been generated. Rather, data have been compared with an existing model and are presented as being in keeping with this model. Thus, by implication, the intervention is seen as being effective.
Summary of results	• All group members wanted to share the techniques they used, the observations they made, the patterns they saw, and the explanations they offered. However, they often failed to achieve intersubjective communication. • These difficulties can be explained by polesemy, privileging and holding the floor. The lack of opportunity for holding the floor of one participant was not a result of overt prejudice in speech and action. The expectations about how and when people should talk, how work should be done, and what standards of quality they should aspire to led them to reconstruct inequities of society as a whole.
Conclusions	There are four areas where improvements could be made in group work (pp 309–310). 1. Talk explicitly with pupils about how to maintain productive and equitable participation in groups. 2. Teachers emphasising to groups that success is not about getting the right answers, but about thoughtful engagement with tasks and with each other.

	3. Assigning particular roles to pupils in groups – 'influencing the floor' 4. Trying to promote 'inclusive leadership' in groups – which the authors recognise as very difficult
Weight of evidence A (trustworthiness in relation to study questions)	**Medium** The difficulty is in knowing exactly what the research questions were. But, to the extent that the data yields detailed and carefully documented evidence on the group discussions and group dynamics, it is trustworthy.
Weight of evidence B (appropriateness of research design and analysis)	**Medium** Reasonable detail of sampling frame, but sampling method not very specific. Comparison within the group is part of the design. Some context variables (class) are described. Triangulation increases trustworthiness of data collection, but no instruments are provided. Explicit but imposed analysis scheme results in medium trustworthiness.
Weight of evidence C (relevance of focus of study to review)	**Medium** Understanding of evidence is tangential. Nature of discussion is explicit independent variable. Four sociological construct are highly appropriate measures. Broad range of discussion reported. Group composition slightly construed (maximum diversity).
Weight of evidence D (overall weight of evidence)	**Medium**

Lajoie SP, Lavigne NC, Guerrera C, Munsie SD (2001) Constructing knowledge in the context of BioWorld. *Instructional Science* 29: 155–186.	
Country of study	Assumed Canada
Details of researchers	Researchers at McGill University, Canada funded by Canadian Sciences and Humanities Research Council and Wisconsin Alumni Research Fund
Name of programme	BioWorld computer program/software
Age of learners	14 to 15
Type of study	Evaluation: researcher-manipulated
Aims of study	To examine students' use of Bioworld Computer learning environment to solve problems related to the digestive system and analyse how the student actions and verbal dialogue were conducted to pinpoint the types of features within BioWorld that were most conducive to learning and scientific reasoning

Summary of study design, including details of sample	Students from two grade 9 biology classes worked in pairs to use the BioWorld program. Classes were of comparable ability level. They were allowed to choose their own partners for the task. The entire sample was used for the first two research questions. Data from six pairs were used for research question 3 (role of teacher guided groups and of researcher guided group). Teacher selected these groups as being equivalent in terms of their previous grades and ability to articulate their understanding. Actual sample: 40 students
Methods used to collect data	• Observation: Audiotapes and videotapes • Computer log of actions and decisions on the BioWorld program
Data-collection instruments, including details of checks on reliability and validity	• Limited details are given; data about the students' choices about the diagnosis and how these changed, about access to virtual tests and other information were 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 RQ 3) were used as benchmarks for indicators of student performance in scientific reasoning.
Methods used to analyse data, including details of checks on reliability and validity	Verbal data was not analysed but used as exemplars to support computer data. Statistical for computer data. • Initial one-way MANOVA test was used to determine if there was a difference between students from the two different classes. • A Pearson correlation was used for the features in terms of the relationship between group and expert actions. • A MANOVA 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-low** For the qualitative aspects
Weight of evidence B (appropriateness of research design and analysis)	**Medium** Some detail of the school, but little for the selection method of the discussion group members. No comparison between groups. Context of SGD described in task, strategies adopted, and quotations. Data collection has reasonable contextual validity and triangulation. Data analysis careful and detailed.
Weight of evidence C (relevance of focus of study to review)	**Medium** Dealing with evidence is the main focus of the intervention. Focus was mostly on the computer cues and scaffolding, not the discussion. Measures (sequence and type of information accessed) are not used to test the nature of discussion (dialogue was used to survey other variables). Reports nature of discussion only indirectly. Naturalistic setting – class in computer laboratories.
Weight of evidence D (overall weight of evidence)	**Medium**

Meyer K and Woodruff E (1997) Consensually driven explanation in science teaching. *Science Education* 81: 173–192.	
Country of study	Canada (Toronto)
Details of researchers	Not stated
Name of programme	The learning approach is called ‘consensually driven explanation in science’.
Age of learners	Grade 7 (Canada), aged about 13 to 14
Type of study	Exploration of relationships: Possibly the relationship between the type of input and the subsequent knowledge construction.
Aims of study	To document the (dis)advantages of a consensually driven explanation strategy in constructing scientific conceptual understanding
Summary of study design, including details of sample	The intervention, spread over two weeks, included a pre-test of conceptual understanding of light, and four sequences of prediction-observation-manipulation-explanation based on practical activities providing unexpected observations. After consolidating conceptual models within the group, conflicting interpretations emerged through whole-class presentations. A question/answer session with an ‘expert’ introduced the scientific concept of light. The authors (rightly) justify a longitudinal design for their interest in strategies for building consensus in the conceptual understanding of light. No limitations of design presented. The role of the researchers is unclear. The sample consists of an intact Grade 7 class of 19 students in a white and suburban school near Toronto. The class was divided into five groups of three or four students, and one of these groups (the only one with three members and a gender mix) was studied. Ability distribution within groups is not provided. Authors state that the focus group was typical of the three other groups. Selection was based on being vocal, and co-operative. No discussion of strength/weakness of data sources.
Methods used to collect data	• Observation: presume audiotapes • Self-completion report or diary: student records (such as worksheets) • Other documentation: records of students’ written work
Data-collection instruments, including details of checks on reliability and validity	• Audio-recorded group conversations, field notes (no format provided) and students records (some detail of prescribed record). • No other detail of data-collection instruments • Multiple sources gives possibility of triangulation. Verbatim quotations provide richness (sometimes obscure).
Methods used to analyse data, including details of checks on reliability and validity	• Matrix of discourse, identifying types of statements (e.g. predictions and justifications), questions and answer to each other; descriptions of observations; explanations; talk around material manipulation; variables being converged on. • Flowchart of discourse for all four experiences. Identify surviving or abandoned ideas in explanations, and common variables and manipulation sequences across the four experiences. The analysis focused on a group’s convergent ideas, advances in understanding

	and the coherency of their explanations. It is unclear how the classification (e.g. predictions and justifications, questions and answer to each other; descriptions of observations; explanations; talk around material manipulation; variables being converged on) and the flow chart categories (e.g. surviving or abandoned ideas in explanations, and common variables and manipulation sequences across the four experiences) have been generated. Probably from common sense. • Contexts portrayed by several quotations of group interactions. Some are difficult to interpret. • Little attention to multiple perspectives, apart from the students' summary reflections on expert visit. Diversity is portrayed in student discourse but the study was seeking *consensus*. • Patterns are not presented, but must have been the essence of the analysis of the flow charts. • No detail on specific trustworthiness of the analysis process. • Findings well projected against other literature. • This study does not intend to evaluate the intervention, and unintended outcomes are not highlighted.
Summary of results	Three mechanisms are determinants in the consensus-building process: • Mutual knowledge: Ddisagreement in initial conceptual understanding sets the task of creating mutually accepted knowledge within the group. This need did not emerge during the prediction phase because of similar knowledge, and the fact that there was no need to justify the prediction. • Convergence: During this discourse process, members try to add to their knowledge. Convergence is particularly evident when observations can not be explained with existing knowledge. Convergence is encouraged by manipulating materials, and by 'What if?' questions. Groups converge usually on one variable only. • Coherency: This mechanism responds to the need to 'fit' an explanation across different phenomena. Coherence is particularly prominent because of the sequence of four experiments, which requires looking for patterns and anomalies across phenomena.
Conclusions	• There are three mechanisms for consensus building in small group discourse: mutual knowledge, convergence, and coherency, and that this forms a framework for consensually driven explanations in science education. • Students need time to adjust to a collaborative enquiry approach that requires them to generate and evaluate their own ideas and to share with others. • Students were uncertain what they had learned as they had not memorised a textbook explanation.
Weight of evidence A (trustworthiness in relation to study questions)	**Medium** The research question is difficult to identify. The study is ambiguous about the role of students' discourse. Sometimes the paper indicates that the discourse itself is the focus of the study, in the analysis the authors see the discourse only as a explicit record for changes in students' conceptual understanding. The description of the group of three students and the class from which they were selected is reasonably complete. The data-collection methods are diverse and provide the opportunity of triangulation. The analysis does not address the nature of the discourse. No attempt has been made to present contrasting incidences. Several of the quotations are not fully commented on, and thus it remains unclear how they have been interpreted. Relationship between the findings and conclusions could be stronger.
Weight of evidence B	**Medium-low**

(appropriateness of research design and analysis)	Nature of sampling frame (school/class) is well described, and the characteristics of the small group too, but not the reasons for selecting this one small group. No comparisons: no relationships between characteristics of group members and the nature of discussions presented. Reasonable context (task/solving strategies), none on interpersonal relationships. Data collection is solid, but focuses mainly on collecting development in conceptual understanding. The analysis disregards the nature of the group discussion.
Weight of evidence C (relevance of focus of study to review)	**Medium-low** The main focus of the intervention is the development of conceptual understanding from evidence provided, not of the understanding of evidence. The discussion is used as a way of explicating conceptual understanding. The nature of group discussion is not studied. Measures are geared mainly to tracing conceptual understanding. No breadth to data. Discussion group is a little artificial (shows gender mix and is composed by teacher), although set within actual classroom.
Weight of evidence D (overall weight of evidence)	**Medium-low**

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
Aim of study	To evaluate the effects of an intervention including guidance of the use of scientific explanations and constructive group interaction on the ability to apply knowledge of kinetic molecular theory to everyday problems
Summary of study design, including details of sample	The collaborative problem-solving programme involved using a sequence of activities on kinetic molecular theory with nine Grade 6 classes in two schools over a period of two years. Students were placed in groups of four, heterogeneous with regard to gender and race. Discussion tasks were aimed at modelling the working of scientific communities. A variety of data were collected (see later sections). This study focuses on analysis of discourse. Actual sample: Nine classes with an average of 26 students implies a sample size of around 230 students.

Methods used to collect data	• Curriculum-based assessment: pencil-and-paper tests of conceptual understanding • One-to-one interview • Observation: video recordings of particular groups • Self-completion report or diary: student logs
Data-collection instruments, including details of checks on reliability and validity	• No details given
Methods used to analyse data, including details of checks on reliability and validity	• The use of a t-test for pre- and post-intervention results is assumed. • Grounded theory seems to have been used for the analysis of group and class discussions. With comparison between year 1 and year 2 observations. • Checks on reliability: No details given, other than, by implication, multiple data sets enhance reliability. • Checks on validity: Triangulation between student logs and recorded group discussions forms some type of validity. Authors do not mention having done this.
Summary of results	• Students initially approach problem-solving very differently from adult scientists, in ways in which teachers would characterise as careless, immature or unthinking. This changed over time. • Poster presentations revealed contradictions in results, which in turn led to discussion of accuracy of reporting. • Students initially found whole class discussion and debate about reaching a consensus confusing, but did ultimately arrive at an agreed scientific view. • Students enjoyed planning the investigation. • Students used explanations to scaffold their discussions, particularly to provide reasons for their proposals. • Students also discussed explanations. • Students stayed focused on discussion tasks. • Students were able to use their previous everyday experience to inform planning of investigations. • Students demonstrated some of the characteristics of engaging in the enterprise and language of science, particularly in the second year of the study. • Post-test measure of understanding showed a significantly greater number of students in year 2 achieved the targeted conceptual goal. • No significant difference in pre-test for year 1 and pre-test for year 2 [t(82) = 1.05, p = 0.296], but significant difference on the post-test [t(82) = 2.625, p = 0.005]. On the post-test 36.6% in year 1, and 51.1% in year 2 provide explanation for dissolving including both macro and micro-elements; 24.4% in year 1 and only 6.4% in year 2 provide naive responses.
Conclusions	• Specific conclusions of the study are not summarised, but are implicit in the reporting of the data. • The conclusions focus on teacher needs to support the use of activities such as those described in the paper.

Weight of evidence A (trustworthiness in relation to study questions)	**Medium** The 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** Sample frame (classes) reasonably described, but no details of discussion groups. Good details of how the groups were composed. In fact, the unit of analysis remains the class. Comparison between consecutive year groups based on difference in task. Reasonable detail of context of SGD (tasks, preparation). Data collection (several in-tact classes) inappropriate for this review question. Data analysis mainly consisted of descriptive assertions with very few quotations of the discussions.
Weight of evidence C (relevance of focus of study to review)	**Medium-low** Understanding of evidence (constructing explanations) is used as guidance for group activity. The focus of the intervention is the ability to apply science concepts to everyday situations. Nature of discussion minor element in independent variable (intervention, nature of task). Measures for analysis of dialogue difficult to identify. Quantitative measures for conceptual knowledge not related to this review. Natural classroom setting.
Weight of evidence D (overall weight of evidence)	**Medium-low**

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	
Country of study	USA
Details of researchers	Not given
Name of programme	Not applicable
Age of learners	Grade 10: 11 to 16
Type of study	The report describes how students discussed the problems and attempted to solve them.
Aims of study	The purpose is to understand (i) how student talk in small groups reflects the process by which students solve scientific problems, (ii) the difficulties students encounter when developing scientific arguments and negotiate their social roles, and (iii) the ways these interactions shaped the arguments themselves.

Summary of study design, including details of sample	Data were collected on four group investigations of progressively increasing complexity (providing non-explicit longitudinal and consistency aspect of design). This staging and timeframe was determined by the teaching programme, not the design. Multiple data sources provided triangulation. There is no overall justification given for the approach taken, although aspects of the analysis are justified (top of p 843). Details of sample: 24 students, one intact class. Researchers were 'interested party' (curriculum designers of intervention, team teachers) and data collectors. Six discussion groups (Gr10) of four mixed ability/gender students in an intact class, all but one Caucasian. There is no information about socio-economic status and language skills. The composition of the groups was modified on request or after observing a lack of co-operation. There is no information on typicality of sample. Selection methods seem influenced by access (previous research experience with teacher). No information on school, stream or class/lab setting. No discussion of strengths/weaknesses of sample/method.
Methods used to collect data	• Observation • Self-completion report or diary • School/college records (e.g. attendance records)
Data-collection instruments, including details of checks on reliability and validity	• Audiotape of all six small groups at all their discussions (60 hours). • Videotape of two to six small groups selected at random during all the discussions. • Videotape of all six groups presenting their findings to the class. • Overall eight hours of videotape • Field notes made after each lesson by researchers about 'significant events' and researchers' interpretations • There were no data collection forms or instruments. • Audiotaping of discussions of each of six groups. Field notes of significant events, reflective ideas and subsequent teaching plans. • Trustworthiness increased by triangulation of video- and audio-recordings, with field notes. Researchers were participant observers, providing insight in group discussions, and opportunity for asking for clarification. At the same time risk of directing group discussions.
Methods used to analyse data, including details of checks on reliability and validity	• Audiotapes: These were transcribed (but no mention of whether transcribed in full). Videotapes: These were looked at, but no detail about how or by whom. Notebooks: These were used to identify 'concepts with which students were struggling and features of their social interactions'. • Search for concepts students were struggling with, and features of the interactions (e.g. frequency, content, intent and consequences of individual's contributions to discussions), and the extent to which contributions were task-related. Task engagement of each student was assessed from the data and quantified. Within- and cross-group comparisons were made. • Argumentation ability was classified according to discussion of (pre-set) stages of investigation. Engagement for design, implementation, and interpretation of laboratory investigations was measured. Indicators (emerging from data) were laying out problem's foundation and tools for solving it, or constructive questioning problem-solving strategies. Social roles were grouped in leaders, helpers, and active and passive non-contributors (from data). Leadership styles was differentiated in inclusive, persuasive and alienating (from lit? not mentioned). • Descriptions of handling apparatus, or writing activity provides context of data sources, but there is no attempt to take into account the characteristics and background of the participants or the setting.

	• Detailed provision of a variety of contrasting verbatim group interactions, with interpretive comments, provides multiple perspectives, comparisons. No search for negative cases. • Patterns in data on argument construction, engagement and leadership styles well presented through quotes. • Independent coding of social roles (inter-coder agreement of 100%) indicates high reliability. Overall the analysis reported is rich and plausible. The researchers give good extracts from the conversations and plenty of detail. • No corroborating evidence from other studies used, but triangulation possible. • No information about criteria for effectiveness or impact, and no reflection on unintended consequences of intervention.
Summary of results	• During the course, many students made considerable progress – levels of engagement rose, and students' arguments became both more sophisticated and better situated in an intellectual context. • Progress depended on group dynamics, depending in turn on the style of the group's leader. Inclusive leadership allowed substantial engagement in depth of discussions and number of participants. Persuasive leadership allowed high engagement of the leader, but engagement of other members was limited to procedural rather than interpretive tasks. Alienating leadership generated a lot of off-task talk and engagement was generally low. • In inclusive groups, most members succeeded in connecting new knowledge to the larger intellectual picture. In the persuasive groups, only the leader generated such connections. Alienating leadership resulted in little concern for such connections. • In the inclusive and persuasive groups, the quality of arguments was high (co-constructed in the first case). The group with alienating leadership had fragile arguments and had trouble substantiating their claims under scrutiny.
Conclusions	• The three goals (engagement, placing new knowledge in intellectual context, construction of argument) can be supported by requiring distributed responsibility during group presentations; completion of individual reports based on group work; development of inclusive leadership and equitable classroom participation.
Weight of evidence A (trustworthiness in relation to study questions)	**Medium-high** Some ambiguity about the research questions. Context of sample clearly explained; group composition (mixed-ability and gender) in line with research questions. Elaborate data-collection methods strengthened through triangulation. Diverse analysis methods: Some indicators emerging form the data, others pre-set from the literature. Relatability could improve by more extensive description of the background of students and context of school.
Weight of evidence B (appropriateness of research design and analysis)	**Medium** No sampling frame. Good comparison between discussion groups, and different stages of investigations. Context of SGD could be more detailed. High trustworthiness of data collection. High trustworthiness of analysis method.
Weight of evidence C (relevance of focus of study to review)	**High** Understanding of evidence (argumentation) one of the foci. Nature of discussion explicit independent variable. Measures (audio-/video-recording with participant observer) highly appropriate for testing nature of SGD. Breadth of nature of group discussion. Highly representative of SGD in classrooms.
Weight of evidence D (overall	**Medium-high**

weight of evidence)	

Roth W, 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	
Country of study	Canada
Details of researchers	Not stated
Name of programme (if applicable)	Not applicable
Age of learners	11 to 16 and 17 to 21
Type of study	Exploration of relationships: linking concept mapping, discourse and concept understanding.
Aims of study	• To describe and analyse: 1) the process of concept mapping, 2) the student-student and student-teacher interactions and 3) the cognitive activity of the participants.
Summary of study design, including details of sample	No particular rationale is given. It is taken as read that the four sources of data (observation, transcript of video recordings, concept maps generated in discussions, written student reflections) are appropriate ways to address the aims of the study. This seems reasonable. The researcher taught all the students participating in the study (explicitly stated). The overall sample consists of 46 and 48 students on the junior level physics course in years 1 and 2 of the study respectively, and 29 and 25 students on the senior level physics course in years 1 and 2 of the study respectively. The 25 senior students in year 2 of the study were all from the group of 46 junior students in year 1 of the study. No other details are provided, other than that the school where the study took place is a private school, but the implication is that they are reasonably representative of the wider student population. Note: All the data ultimately presented focus on one group of three male students of varying ability. The sample is opportunistic in that it is simply the classes taught by one of the researchers. No information is given about strengths and weaknesses of data sources and methods.
Methods used to collect data	• Observation: Including transcripts of video-recordings • Self-completion questionnaire • Other documentation: Concept maps generated in discussions • Not clear how students' reflections were collected
Data-collection instruments, including	• Student group and class activities videoed and audio transcribed. Student concept maps collected and studied. No other details are

details of checks on reliability and validity	provided about data-collection methods. • No details of observation schedules (if used) are provided. No details are given of any instrument used to gather student reflections. From the description of methodology (p 536), it is unlikely that any structured observation was done. • By implication, multiple data sources increase trustworthiness.
Methods used to analyse data, including details of checks on reliability and validity	• The researchers say they adopted a technique used by anthropologists studying interactive behaviours. Both researchers watched the videos and read the transcripts to form tentative descriptions. These were refined, modified or discarded on basis of further comparisons within sets of data collected. Disagreements were discussed until a consensus was reached or discarded if no consensus was reached. Categories from the data were used to characterise the interactions between participants and the concept maps they produced. (In this context, 'participants' included the teacher.) A robust procedure is described that agrees descriptive categories or discards descriptive categories that cannot be agreed between the researchers. No specific detail of validity, but analysis of discourse was checked against concept maps. • No information is given about context of data sources. • No examples are included to illustrate diversity in the data. • No information is given about divergent positions in the data. • Evidence is taken from different the data sources in the study. • The findings are related to a range of other work (p 548 onwards) • Criteria for impact are not relevant to this study. • No information is given on unintended consequences.
Summary of results	• Group discourse over the construction of a concept map provides a vehicle for negotiation of meaning and understanding of concepts and their relationships. • Discussant positions are stated, contested and views either accepted or temporarily or permanently rejected. Temporarily rejected positions can become accepted. Positional finally stabilise to express taken-to-be-shared meaning as a map is constructed. • Students can form strategic alliances in support in support of a position. A position is seen to have more weight if the discussant is known to have has a special interest in the area. • There were examples of collaboratively constructed concepts, adversarial exchanges and temporary alliances as the concept map was constructed. • Agreement on a position was often reached with reference to authority, to a majority view or to a common lower order of agreement. This agreement was not always based on a common understanding. • The process of construction of a map was contingent on specific local conditions. • The concept maps became a tool for negotiating meaning. • Mapping concepts as a group activity may be more important than the concept map itself. • Students tended not to engage very often in processes which foster meaning. Rather they would reach agreement on the basis of finding something agreeable to all group members, authority as the deciding factor, and majority rule. • Moveable paper slips for developing the maps worked better than drawing maps as they reduced the tendency to produce a product with the least amount of redrawing.

Conclusions	• Concept mapping provided a framework in which students engaged in sustained discourse over periods of an hour's length. • The fixed set of concepts delimited the content of the discourse. • The students not only linked pairs of concepts, but built a map of a thematic territory. • Taken together, this means concept mapping provides a structure through which students can learn the language patterns of science and with it, construct scientific knowledge. • Students struggled with language, often making short utterances, and did not clarify their understanding because they did not resort to explanations, justifications and elaborations. • A major outcome of the study is the recognition of the need to help students to argue and to use evidence to support a proposition.
Weight of evidence A (trustworthiness in relation to study questions)	**Medium** The conclusions have been drawn from more evidence than is presented in the paper, and the analysis is careful and detailed.
Weight of evidence B (appropriateness of research design and analysis)	**Medium-high** Opportunistic sample, and little detail given. Comparisons made within group only. Highly detailed description of study. Data-collection methods appear to be very trustworthy. Analysis methods robust.
Weight of evidence C (relevance of focus of study to review)	**High** Understanding of evidence is main focus of intervention. Nature of discussion is explicit independent variable. Measure highly appropriate for testing nature of small-group discussions. Range of discussion is reported, although from a small sample. Naturalistic setting.
Weight of evidence D (overall weight of evidence)	**Medium-high**

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, China
Details of researchers	University-based researcher working on funded project. A research assistant is also mentioned. There are some indications in the text to suggest elements of practitioner research or research undertaken for a higher degree, although no details are given.
Name of programme	No details given

Age of learners	17 to 18
Type of study	Evaluation: naturally-occurring
Aim of study	To explore whether and how group discussion of feedback of multiple alternative solutions to qualitative physics problems helped to improve students' problem-solving skills and understanding of underlying physics concepts
Summary of study design, including details of sample	A case study focusing on the evaluation of three qualitative physics problems The sample consisted of a convenience sample of one class of 18 Year 12 students, of whom 16 were included in the analysis. The study involved four stages: a pre-test, feedback, a post-test (of three parallel questions similar to the three in the pre-test) and semi-structured interview. In the first two stages, students worked in dyads, and their peer-interactions were audio-recorded. The post-test and interview involved individual students.
Methods used to collect data	• Curriculum-based assessment (physics problems) • Group interview • One-to-one interview • Audiotapes 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 to study questions)	**Medium-low** Indicators for problem-solving skills are not clearly stated. Reported abilities (e.g. meta-cognition) are unrelated. Reliability and validity of data-collection methods and analysis methods not specified. The validity and reliability of data-collection method and analysis method is high for RQ 2. The research design could have included a control group.
Weight of evidence B (appropriateness of research design and analysis)	**Medium-low** Reasonable detail on the class, some about the school, nothing about the participants in individual discussion groups. No comparison, but only description of alternative solution strategies. Tasks and strategies provided in detail. No information is given about the context of group relationships. Data collection is mainly based on pre- and post-tests which is not relevant to this review question. Data analysis of qualitative data focuses on the understanding of the physics content.
Weight of evidence C (relevance of focus of study to review)	**Medium-low** The focus of the intervention is on the effect of a presentation of several correct solutions to a problem on problem-solving skills, not on understanding of evidence. The discussions are a secondary discreet element of the study; the main focus is problem-solving skills. Measures of discussion (expressing different concepts, metacognition) do not test the nature of discussion. No breadth. Highly motivated classroom less 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 students 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 students 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 reliability and validity	• 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% 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	• Mean scores for each student 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

validity	• 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 explanations offered by 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 leads to discussion of factors at work, and taking these into account by reconstructing their explanations. • Female pairs learn by identifying but ignoring differences in predictions and feedback. Although 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 co-ordination of ideas and evidence (as in all-male), and no co-ordination between ideas relevant to different problems (as in all-female).
Conclusions	• 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, but rather, to adapt to the apparent requirements of the female and mixed pairs which were weak at predictions. Suggestions are made by the authors of ways that could assist predictive discussion: for example, 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 research design and	**Medium-high** Some information on the school context, considerable information on sampling method, but little else on the sampled discussion groups, students or selection justification. Distinct comparison of between and within gendered groups in the design. Contexts of task and

analysis)	strategies detailed, but little about context of group interaction. Data collection high trustworthiness with accommodation of various gender combinations, over age range. Data analysis (using established classification scheme) of dialogue structure not dialogue itself. No quotations.
Weight of evidence C (relevance of focus of study to review)	**Medium** The main focus of the intervention was on dealing with self-generated evidence. The study looked explicitly at the nature of the dialogue. Measures classified structures of dialogue, not nature of dialogue itself. Breadth of measures (explanation, prediction) reasonable. Unrepresentative for classroom situation (artificial groups, across 12 to15 age range).
Weight of evidence D (overall weight of evidence)	**Medium-high**

Tsai 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	
Country of study	Taiwan
Details of researchers	Main researcher was university-based. Eight researchers did the observation
Name of programme (if applicable)	Not applicable
Age of learners	Grade 8: aged 13 to 14, from two classes in a junior high school in Taiwan
Type of study	Exploration of relationships: (1) scientific epistemological views and peer interaction in lab work; (2) scientific epistemological views and perception of actual learning environment; and (3) scientific epistemological views and preferred learning environment.
Aims of study	To investigate: (1) To what extent, and in what way, are there relationships between students' scientific epistemological views (SEVs) and their social verbal interactions in lab activities? (2) To what extent, and in what way, are there relationships between students SEVs and their perceptions about actual and preferred learning environments? (3) How do student interview results substantiate quantitative findings and then help interpret the interplay between students' SEVs and their views about the nature and the aims, values and other relevant beliefs of laboratory activities?
Summary of study design, including details of sample	• A pre-existing instrument was used to assess students' SEVs (Pomeroy's questionnaire). Eight trained researchers each observed one or two subjects in each class and recorded categories of students' oral peer interactions over six lab sessions. A pre-existing instrument (the Science Laboratory Environment Inventory, SLEI) was used to explore students' perceptions of laboratory activities. Two sessions

	were videotaped to validate observation data. Twenty-five students were interviewed to add depth to the data. No explicit rationale is given for the study design, but the implication is that the range of instruments and techniques used to gather the data provides a detailed picture. • The status and role of the researcher is not made explicit, but was not involved in teaching, observation or interviewing. • By implication, the sample is taken to be a reasonably typical sample of junior high school students in Taiwan. • No rationale is given for selecting the schools and classes. By implication, the sample is opportunistic. However, within this, the observations were focused on students who expressed a strong certainty or confidence about their SEVs based on their responses on the instrument used to assess SEVs. This is justified by saying these students were expected to be highly aware of their epistemological orientations towards science (top of p 657), an important variable in the study. • No information is given about strengths and weaknesses of data sources and methods. • Details of sample: 86 students did the SEV instrument. 28 students were identified as expressing confidence in the SEVs; 25 students took part in the remaining phases (3 of the 28 were absent). Students worked in groups of five or six.
Methods used to collect data	• Curriculum-based assessment • One to one interview (face to face or by phone) • Observation • Self-completion questionnaire • Exams: measures of achievement taken from scores on school-wide science tests
Data-collection instruments, including details of checks on reliability and validity	• A pre-existing instrument was used to assess students' SEVs (Pomeroy's questionnaire). Eight trained researchers each observed one or two subjects in each class and recorded categories of students' oral peer interactions over six lab sessions. • A pre-existing instrument (the Science Laboratory Environment Inventory, SLEI) was used to explore students' perceptions of laboratory activities. Two sessions were videotaped to validate observation data. Twenty-five students were interviewed to add depth to the data. No explicit rationale is given for the study design, but the implication is that the range of instruments and techniques used to gather the data provides a detailed picture. These data were used to measure aspects of the sample as findings of the study. Note: No examples of items from SEV and SLEI instruments given, and no clear picture emerged of exactly what was being noted down in the observations. • Checks on reliability: The use of pre-existing instruments seems to be taken as an indicator of reliability and validity. Interview transcripts were done after translation from Chinese to English. An independent Chinese speaker checked the accuracy of the translation. • Checks on validity: Established instruments used. Translation to and from Chinese checked. • By implication, the use of multiple data sources increases trustworthiness. Trustworthiness was also increased by the use of established instruments, the use of trained researchers, the discarding of first observations, researchers observing different students in different observation sessions.
Methods used to analyse data, including details of checks on reliability and validity	• Students in sample are those who were confident about their SEV scores. Discourse segments were analysed with respect to Shepardson's five major negotiation categories: negotiation of status, negotiation of action, negotiation of meaning, negotiation of materials and other. However, it is not clear if this was done at the time of observation and formed the record of the observation or done after the lesson from some record of the lesson. The two pre-existing instruments appeared to have their own analysis schedules. Interviews were translated from Chinese and then transcribed. No other details of analysis are provided. Correlation coefficients were

	calculated between verbal negotiations and a range of other variables: verbal negotiation activities, perceptions of laboratory environments, perceptions of preferred laboratory environments. • No information is given about context of data sources. • Multiple data sources: questionnaires/inventories, observation, interviews, test results • No information is given about divergent positions in data. • No details are given of reliability or validity. Multiple data sources increase trustworthiness. • No corroborating evidence is used. • Impact is not an issue in this study. • No reflections on unintended consequences are included.
Summary of results	• High achievers tended to have more verbal interactions directly related to laboratory activities. • Students having SEVs more oriented to constructivist views of science tended to negotiate the meanings of laboratory activities with their peers and teachers than those with empirically-aligned SEVs. • Constructivist learners tended to believe that school laboratory environments did not emphasize open-ended approaches to investigations, or integrate with theory classes, and viewed them less positively than other students. • Constructivist students favoured laboratory environments where students were supportive of each other and activities highlighted an open-ended approach, and were frustrated by traditional laboratory work. • Empiricist students placed greater emphasis on doing lab work following codified procedures of science texts and they believed that lab exercises made science concepts more impressive, acting as memory aids. • Interview data tended to support these conclusions.
Conclusions	• SEVs are related to learning in laboratory activities. Traditional laboratory experiences can be frustrating for students with constructivist-oriented SEVs, although they tend to develop better understanding. These students prefer an open-ended, peer-negotiated approach. • Science teachers should carefully consider students' epistemological views of science when planning lab work especially with regard to creating a peer-supported atmosphere and emphasising an open ended manner of experimentation for constructivist students (see p 670) • From the abstract: "An appropriate understanding of the constructivist epistemology of science should be an essential prerequisite for implementing so-called constructivist teaching."
Weight of evidence A (trustworthiness in relation to study questions)	**Medium** The study focuses on a particular subset of students and it is difficult to generalise from this to answer the RQs posed in other than a very narrow sense.
Weight of evidence B (appropriateness of research design and analysis)	**Medium-low** Reasonable detail of sample. No comparison between or within groups. Little contextual details provided. Reasonable trustworthiness of data collection. Medium trustworthiness of analysis.

Weight of evidence C (relevance of focus of study to review)	**Medium-low** Understanding of evidence is tangential to intervention. Nature of discussion is major discrete element of study. Measure only indirectly appropriate for testing nature of discussion. Nature of discussion reported only indirectly. Selectivity in sample means there are limits to how representative study is of SGD is classrooms.
Weight of evidence D (overall weight of evidence)	**Medium-low**

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.	
Country of study	Not stated but assumed to be Canada (both researchers are based in Canada).
Details of researchers	Meyer and Woodruff (1997) report their own field notes as a source of data and so must have been involved in that study. Their involvement in this study is not clear but one data set seems to be the same as that in the previous study. Teacher made field notes. Teacher acted as participant observer.
Name of programme (if applicable)	Not applicable
Age of learners	5 to 10 and 11 to 16: Most pupils are in the 11 to 13 age range (US/Canadian grades 5 to 7).
Type of study	Exploration of relationships
Aims of study	To explore the patterns of inter- and intra-group discourse as middle students explain particular phenomena, integrating the findings from three studies. • The discussion focuses on three perspectives: (1) inquiry and the generation of explanations in small groups of students, and the validation of explanations in large groups of students; (2) the evolution of student explanations within inquiry discourse; and (3) pedagogical interpretations and considerations.
Summary of study design, including details of sample	Study involved tracking the intra- and inter-group discourse in three classes as they studied topics on shadows and images (Grade 7) and floating and sinking (Grades 5 and 7). It is taken as read that the design of the study is appropriate. However, other than a mention of one class teacher involved in the study acting as a participant-observer, little detail is provided of the study design and methods. Reference is made to small groups of three to four students, but, other than work being done with whole classes, no reference is made to sample size. The roles of the researchers are not clear.

	No characteristics of the sample are described. The sample size is opportunistic. No information is given about the strengths and weaknesses of the sources and methods.
Methods used to collect data	• Observation: Reference is made to teacher acting as participant observer. The nature of the extracts implies they were tape recorded and transcribed. • Other documentation: Field notes taken by teacher
Data-collection instruments, including details of checks on reliability and validity	No information given.
Methods used to analyse data, including details of checks on reliability and validity	• While the paper lacks detail of data analysis and the construction of argument and counter argument, the researchers propose three stages in the evolution of students' explanations of science phenomena: (1) explanations focus on descriptions of properties rather than mechanisms or relations; (2) explanations focus on descriptions of a set of relations among variables students believe to be relevant to the context; and (3) explanations focus on relative conditions and involve a complex system of priorities applied to conditions that attempt to explain mechanisms that can account for a range of related phenomena. Note: Meyer and Woodruff (1997) refer to setting up a matrix of discourse, written explanation and actions for 'effect' (= observed phenomenon) and then using this to generate a flowchart that documented students' reasoning. They then looked for 'explanatory power' that survived, ideas that were abandoned, common variables and manipulations across 'effects'. In general, the analysis focused on a group's convergent ideas, advances in understanding and the coherency of their explanations. • No information is given about context of data sources. • No information is given about diversity in data. • No information is given about divergent positions. • No information is given about measures to increase trustworthiness of analysis. • Data are related to other research studies with similar findings. • Impact is not an issue in this study. • No information is given about unintended consequences.
Summary of results	There are three types of explanations. • First order explanations: expected outcomes – explanations based on common observations and inferences about material objects within the phenomenon, but without causal justifications. • Second order explanations: convergence on variables – deconstructing what matters in a particular phenomenon, or asking and answering 'what if?' questions; students explain phenomena as a context, and explanations typically describe a set of relations among variables believed to be of relevance to the context. • Third order explanations: coherence of related phenomenon – explanations which focus on relative conditions and test the coherence of an explanation with related phenomena • Small-group and whole class discussion helps students refine their explanations.

Conclusions	Students need to become engaged with enquiry and explanations in their science lessons through small-group and inter-group discussions. Such activities motivate students to increase explanatory coherence by abandoning some ideas and advancing others, with students' explanations showing a progressive shift in this process from explanations that focus on properties of objects to dynamic ones that incorporate complex priority systems and relative conditions. Hence such discussions promote students' understanding of ideas, rather than just the accumulation of fact. This requires a change in classroom culture.
Weight of evidence A (trustworthiness in relation to study questions)	**Medium-low** Little information is provided in the report itself about methods and analysis.
Weight of evidence B (appropriateness of research design and analysis)	**Medium-low** No detail is given of sample. No comparisons are made between or within groups. Little detail of context is provided. Data-collection methods appear reasonably trustworthy. Data-analysis methods appear reasonably trustworthy.
Weight of evidence C (relevance of focus of study to review)	**Medium-high** Understanding of evidence is the main focus of the intervention. The nature of the discussion is an explicit independent variable. The measure appear reasonably appropriate for testing the nature of the discussion. A very narrow range of discussion is presented. The situation is naturalistic.
Weight of evidence D (overall weight of evidence)	**Medium**

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	Grade 9: 13 to 14
Type of study	Evaluation: researcher-manipulated

Aim 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). Students worked in group sizes of five to seven. The assignment of classes to experimental and control groups was random. The experimental group received the Genetic Revolution unit, which took twelve lesson of teaching time. It is not immediately clear how many teachers were involved. The implication is eight, of which three taught both a control and an experimental group. Each group received a pre- and post-test of argumentation skills and biological knowledge. A multiple-choice test, audiotaped discussions and written worksheets were used to gather data. Actual sample: Not all students were included in the analysis, due to absence when some of the data were collected. No details of the final samples size are given.
Methods used to collect data	• Curriculum-based assessment: 20 multiple-choice items • Student worksheets • Audiotapes of four small-group discussions
Data-collection instruments, including details of checks on reliability and validity	• 20 multiple-choice items to assess biological knowledge • Worksheets to assess argumentation skills • Audiotapes of four small-group discussions • Checks on reliability: No details about reliability given • Checks on validity: Some of the multiple-choice items were from previous years' examinations and some developed for the study, with the content validity of the latter items being checked by an expert.
Methods used to analyse data, including details of checks on reliability and validity	• Qualitative categories based on previous research were used in analysis of audiotaped discussions. • Researcher-developed method to score pre- and post-tests of argumentation skills • Calculation of inter-rater reliability scores for argumentation analysis • t-test of significance of use of biological knowledge in post-test • t-test of significance of mean scores on argumentation tests • Test of 'frequency of conclusions' • Checks on reliability: Argumentation skills analysis was done by both researchers, and inter-rater reliability scores calculated. • Checks on validity: No details are given.
Summary of results	• Following instruction, the number of students using correct, specific biological knowledge in constructing arguments increased from 16.2% to 53.2%. • Students in the experimental group scored significantly higher than students in the control group in a test of genetics knowledge. • Analysis of the written tasks showed an increase in the number of justifications and in the complexity of argument. • Students were able to transfer reasoning abilities tools in the context of bioethical dilemmas to the context of dilemmas taken from everyday life.

	• There were dramatic changes in the quality of student arguments. • Changes were detected in the frequency of explicit conclusions, the mean number of justifications for a conclusion and in the number of ideas students expressed while talking. • Integrating explicit teaching of argumentation into the teaching of dilemmas in human genetics enhances performance in both biological knowledge and argumentation.
Conclusions	• Students showed improved understanding of biological concepts. • Teaching through social issues provides 'anchored instruction' for students by generating interest and connecting to out-of-school life experiences. • Student learning was aided by having students work in small groups for substantial amount of time in most lessons. • Argumentation skills were enhanced by explicit instruction about the formal structure of an argument, and the generation of multiple opportunities for students to take part in discussions that require intensive use of arguments. • Reasoning about dilemmas should be integrated into other science topics. • The authors advise caution against making unsupported generalisations from their findings as they suggests that many may relate to specific properties of the context of the intervention. They also note that many of the teachers and students were very enthusiastic about the programme, again suggesting caution against over generalising from the findings.
Weight of evidence A (trustworthiness in relation to study questions)	**Medium** Possible researcher and teacher bias mean that the findings have to be treated with some caution. No details are given of how schools and teachers were recruited into the study.
Weight of evidence B (appropriateness of research design and analysis)	**Medium** RQs 4 and 5 relevant to this review. Good detail of sampling frame (school, teachers) but none for the sampling method of the discussion groups. No detail for the nature of the group members. Experiment/control group design (with and without intervention developing argumentation skills). Some context of SGD well described (tasks and discussion strategies, less so on group members relationships). Data collection appeared trustworthy. Use of existing classification scheme for argument structure, some illustrative quotes. Not related to member characteristics.
Weight of evidence C (relevance of focus of study to review)	**High** The intervention focuses on both argumentation skills and conceptual understanding. The nature of discussion (here argument) is explicit variable of the study. Measures highly appropriate but largely based on written work. Good breadth of measures (argumentation, explanation). Typical classroom situation.
Weight of evidence D (overall weight of evidence)	**Medium-high**