TECHNICAL REPORT

ICT in science teaching

The effect of ICT teaching activities in science lessons on students' understanding of science ideas

Review conducted by the Review Group for Science

Technical report written by Sylvia Hogarth, Judith Bennett, Fred Lubben, Bob Campbell, Alison Robinson (University of York)

EPPI-Centre
Social Science Research Unit
Institute of Education
University of London

Report no. 1412T · November 2006

EPPI CENTRE
SSRU Social Science Research Unit
IOE LONDON Leading education and social research Institute of Education University of London

The results of this systematic review are available in three formats:

SUMMARY	Explains the purpose of the review and the main messages from the research evidence
TECHNICAL REPORT	Includes the background, main findings, and full technical details of the review
DATABASES	Access to codings describing each research study included in the review

These can be downloaded or accessed at **http://eppi.ioe.ac.uk**

The EPPI-Centre reference number for this report is 1412T.

This report should be cited as:

Hogarth S, Bennett J, Lubben F, Campbell B, Robinson A (2006) ICT in Science Teaching. Technical Report. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

CONTENTS

List of abbreviations

Becta	British Educational Communications and Technology Agency; formerly NCET (see below)
BEI	British Education Index
CAI	Computer-assisted instruction
CAL	Computer-assisted learning
CD-Rom	Compact disc, read-only memory
CPD	Continuing professional development
DfEE	Department for Education and Employment
DfES	Department for Education and Skills (England and Wales); formerly Department for Education and Employment (DfEE)
EPPI-Centre	Evidence for Policy and Practice Information and Co-ordinating Centre
ERIC	Educational Resources Information Center
ICT	Information and communications technology
ILS	Integrated learning systems
ITT	Initial teacher training
MARS	Model-based analysis and reasoning in science
NCET	National Council for Educational Technology; subsequently Becta (see above)
NOF	New Opportunities Fund
OfSTED	Office for Standards in Education (UK)
PGCE	Post Graduate Certificate in Education
QCA	The Qualifications and Curriculum Authority (England and Wales)
RCT	Randomised controlled trial
REEL	Research Evidence in Education Library
SSCI	Social Sciences Citation Index
TDA	Training Development Agency for Schools (England and Wales); formerly the Teacher Training Agency (TTA)
UK	United Kingdom
USA	United States of America
WoE	Weight of evidence
WWW	Worldwide web

Summary

Background

This review has been carried out on behalf of the Training and Development Agency for Schools which has identified a number of key areas in which systematic reviews of the research literature are desirable. One of these is the effectiveness of Information and communications technology (ICT) applications in teaching and learning in the core curriculum subjects of English, Science and Mathematics.

The issue of the effectiveness and impact of ICT in the core curriculum subjects is important. In science, ICT has opened up a whole range of potential applications. At the same time, a wide range of potential benefits resulting from the use of ICT has been claimed for both students and teachers by a number of groups (policy-makers, researchers, some teachers, employers).

Although there is a significant literature on ICT in science education, much of it takes the form of articles on applications for use in teaching situations: the emphasis is on how to use ICT, rather than exploring its effects. There is a sense in which it is taken rather for granted that ICT is a 'good thing', with students being motivated when they use it, and this leads to better learning. Thus a central purpose of this review is to assess the strength of the evidence base to support the notion that the use of ICT activities in science lessons enhances students' understanding of science ideas.

Aims

The review aims to assess the impact of the use of ICT on students' understanding of science.

Review questions

The main review question is as follows:

What is the effect of using ICT teaching activities in science lessons on students' understanding of science?

The term *understanding of science* encompasses scientific knowledge and explanations (facts, laws, theories), the scientific approach (evidence, scientific methods, prediction, problem-solving and so on) and ideas about science (its limitations, the scientific community, risk and so on). Attitudes to science were not included in the review.

Certain criteria were applied to the studies in the systematic map in order to ensure that only potentially good quality and appropriate studies were included in the in-depth review. These were as follows:

- the type of ICT activity most frequently evaluated in the research reports
- the most appropriate study designs for an evaluation (researcher manipulated/controlled)
- studies involving a pre-post design
- studies involving representative or average/typical students

This gave the following in-depth review question:

What evidence is there from controlled trials of the effects of simulations on the understanding of science ideas demonstrated by students aged 11-16?

The term 'simulation' is used in somewhat different ways in different studies. For the purposes of this review, it was understood in two ways:

- the use of the computers to imitate particular experiments
- the use of computer and other resources to imitate a wider situation and thus provide a virtual environment with a range of facilities

The term was not used in the sense of mathematical modelling.

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 in or excluded from the review, searching (through electronic databases) for studies which appear to meet these criteria, and then screening the studies to see if they meet the inclusion criteria
- *Keywording and generating the systematic map:* coding each of the included studies against a pre-agreed list of characteristics which is then used to generate a systematic map of the area where studies are grouped according to their chief characteristics
- *In-depth review and data-extraction:* summarising and evaluating the contents of studies according to pre-agreed categories
- *Synthesis:* providing an overview of the quality and relevance across the studies in the in-depth review and compiling the weighted findings of the collective studies

Results

The studies identified through the searching and screening process established that ICT was being used to teach science education in a variety of ways. The focus of the studies was very largely on teaching scientific understanding and scientific approach. Very little research was carried out on the applications of science, or on the use of ICT for stimulating ideas about science, such as its limitations or risk. A number of the studies were interested in other aspects: for example, attitudes to science, but these were not included in this review.

Thirty-seven studies met the inclusion criteria developed for the overall research review. These studies were keyworded and formed the basis of the systematic map. The map revealed a number of characteristics of the use of ICT in science education:

- The majority of the studies reported work that has taken place in the USA and Taiwan.
- A little over one-third of the studies concerned Biology topics and just under one-third concerned Physics topics. Very little research has been done in relation to Chemistry education.
- Few authors gave explicit details of the ability range of their participant students. (It was therefore assumed that students were mixed ability or average for their age, unless otherwise stated.)
- In one-third of the studies, students worked individually with the ICT and in eight studies (22%) students worked in pairs. Two-fifths of authors (15 studies) did not give details of how the students interacted with the computers.
- Close to 90% of the studies focused on the students' understanding in respect of scientific knowledge/explanations and one-half on scientific approach; 12 studies investigated both. This interest was spread across Earth Science, Biology and Physics.
- Types of ICT used varied, but half were referred to as simulations, either of experiments or of virtual environments. Virtual environments included a range of other ICT activities and non-ICT resources, and could be defined as 'multimedia'. Thus there is some overlap and flexibility in how the various forms of ICT are described or named.
- Fifty percent of the studies were carried out in one school with several classes. Only four studies (11%) involved large samples over several schools. Nine studies did not give full details of how many schools or classes were involved, although they all gave student numbers.
- Three-quarters of the studies used pre-post testing and half used questionnaires. Test results (that is, post- but no pre-test) were used in a quarter of the studies, as were interviews. Eight studies (22%) observed the student activities.
- Three-quarters of the studies were published in academic journals, seven (19%) as conference papers and one in a book chapter.
- Simulations/virtual environments, multimedia and moving images were used in the same or similar proportions to teach both scientific knowledge and scientific approach. (This is not too surprising given the overlap in these three ICT categories.) Data-logging, databases, the internet and tutorial applications were used more often to teach scientific approach than scientific knowledge.

The ICT activity most frequently evaluated proved to be simulations (19 of the 37 studies).

Nine studies met the criteria for the in-depth review. Following data-extraction and the application of the weight of evidence criteria, seven studies were identified as being of sufficient standard to use in the synthesis; one was rated as medium high and six as medium.

The overall findings are listed below. However, as the sample size was small, the simulations variable, the learning objectives diverse and some of the following observations are based on only one study, a number of the findings should not be used for generalisations.

1. Students' use of ICT simulations helped to improve their understanding of science ideas significantly more effectively compared with their use of non-ICT teaching activities (based on six studies).

2. Students' significantly better understanding of science ideas when using ICT simulations versus their use of traditional (non-ICT) activities can lead to understanding of science knowledge (based on seven studies) and to understanding of scientific approach (three studies).

3. The simulations fell into two main categories: (i) simulation of specific experiments and (ii) simulations of a wider scientific situation, commonly known as 'virtual environments', which could include experimental simulations.

4. The positive effect of students' use of ICT simulations on their understanding of science ideas is independent of the type of simulation, that is, simulations as virtual experiments (four studies) or simulations of a virtual environment (three studies).

5. Students' use of ICT simulations was more effective than using non-ICT teaching activities for supporting *basic* science ideas (from three studies), including the improvement of:

- Bloom's lower levels of understanding (two studies)
- understanding of basis aspects of the scientific approach (one study)
- science knowledge of less advanced reasoners (one study).

6. The improvements in higher understanding (for example, application) of more advanced aspects of the scientific approach (for example, the design of an experiment) and for more advanced (formal) reasoners can be achieved to the same extent with or without simulations.

7. The gains from the students' use of ICT simulations were even further increased when teachers actively scaffolded or guided students through the ICT simulations (two studies). The extra gains resulting from teacher guidance through the ICT simulation included further improvement of lower levels of understanding of science (knowledge) and of the scientific approach, including the application of science knowledge to new situations (two studies).

Thus simulations can bring benefits to students in respect of scientific knowledge/explanations and approach, but not in all situations and with all students and teachers. Care needs to be taken in establishing the particular benefits for particular learners and learning objectives in particular situations.

Conclusions

Strengths of the review

The review has a number of strengths:

- The focus is one that is very relevant to the increased use of ICT in science teaching and learning. In particular, simulations are shown to be used in a wide range of situations. Evidence for this comes from the review map, in which 19 of the 37 studies (51%) have simulations as their core mode of ICT.
- The evaluation studies considered student achievement in the spheres of scientific understanding and scientific approach.
- The approach to the review set high standards for the in-depth sample as only evaluation studies that had a control and pre-post test design were included. Additionally, the review only involved those studies that ensured their measures and their methods of analysis were valid and reliable.
- Quality-assurance agreements are high for all stages of the review.

Limitations of the review

There are four main limitations:

- Although 19 evaluation studies involving simulations were found for this review, only seven were of a sufficient standard to include for the synthesis. These can thus only present successful examples of possibilities for teaching and learning in science education and highlight pedagogical points for consideration when using simulations; generalisations cannot be made.
- Some of the terms used in the field of ICT and education appear to be rather fluid. Thus model/modelling/a model can be used in the sense of 'to mimic or represent' or could mean to provide a predictive facility or process. Similarly, simulation can be used to mean that something has been modelled. In this review, the predictive and more mathematical use of modelling was not included. (It did not feature as a topic for evaluation studies.) Multimedia can also include

simulations, in which case it is necessary to tease out the particular contribution of the simulation to learning effect.

- The in-depth studies covered the subjects of Earth Science, Biology and Physics. Only two studies of the 37 in the map and none in the in-depth sample were in Chemistry education. There appears to be a gap in this area of research which has impacted on this review.

- None of the in-depth studies was carried out in UK schools and thus the findings might not be directly applicable to the British educational system. However, the fact that similar findings did emerge from three different countries (USA, Taiwan and Israel) does suggest that there is a measure of robustness in the findings that would make them of use in the United Kingdom.

Implications for policy

Evaluation studies have found that ICT, and simulation in particular, can be helpful in teaching science understanding in respect of both scientific knowledge and scientific approach. However, it should be noted that there is a scarcity of high quality research in the area in which the in-depth study focused.

Teachers will also need training in the use of the simulations to obtain the greater benefit for student understanding. In particular, this review has shown that the use of ICT simulation needs to be carefully integrated into the teaching and learning process, and informed guidance provided. This guidance may be built into the software so that the students may work semi-independently, or it may be provided by the teacher. However, teacher guidance is the more effective. This has implications for policies for initial teacher training and continuing professional development (CPD).

Implications for practice

The review has indicated that there is a lack of clarity in the way that ICT and especially simulations, models and multimedia are interpreted. One implication for practice is that teachers should be made aware of this.

The development of ICT simulations for a large variety of virtual experiments and virtual environments would provide a number of teaching and learning benefits. These include, *inter alia*, saving experimental time and resources, reducing the need to kill animals for dissection, allowing students to repeat experiments with ease, and providing experiences (through virtual environments) that would not otherwise be available to students.

The importance of the structured or guided use of ICT in particular simulations needs to be stressed to teachers. It is not sufficient just to provide the software, unless it has in-built guidance or a virtual mentor. Without either of these, the teacher needs to provide that support. Teachers may also need induction or training if the simulation is part of a complex teaching programme.

The inclusion of simulation activities within science Post Graduate Certificate in Education (PGCE) programmes would also encourage their use.

The newly established Regional Science Learning Centre could provide ideal opportunities for CPD in the use of ICT in science education.

Implications for research

The low numbers of high quality research studies into the value of using ICT in science education, especially in Chemistry, was surprising given the potential benefits. The use of ICT is likely to increase rather than decrease in schools in the near future. It is also likely that curriculum developers and commercial enterprises will increasingly develop software packages for science education. It would therefore be of significant advantage if any science education ICT, of whatever origin, is carefully evaluated before it is adopted.

CHAPTER ONE
Background

1.1 Aims and rationale for current review

The Training and Development Agency for Schools has identified a number of key areas in which systematic reviews of the research literature should be carried out over a three-year period from the autumn of 2003 to the autumn of 2006. One of these is the effectiveness of information and communications technology (ICT) applications in teaching and learning in the core curriculum subjects of English, Science and Mathematics. The present review looks at the second of these subjects, Science. This follows a review in English in 2003–4; a review in Mathematics is taking place at time of writing.

The issue of the effectiveness and impact of ICT in the core curriculum subjects is important. In science, ICT has opened up a whole range of potential applications, including the following:

- practising problems through 'drill and skill'
- providing tutorial instruction
- making use of integrated learning systems
- making use of simulations
- modelling
- using databases and spreadsheets
- data-logging
- controlling and monitoring experiments
- graphing
- working with interactive multimedia (for example, CD ROMs)
- accessing information from the internet
- presenting and communicating information

A wide range of potential benefits resulting from the use of ICT has been claimed for both students and teachers by a number of groups (policy-makers, researchers, some teachers, employers). For example, in the UK, a report by a government body, the National Council for Educational Technology (NCET, 1994), listed well over twenty such benefits, including:

- making students' learning more effective
- increasing students' motivation
- enhancing students' sense of achievement
- providing students with access to richer sources of data and information
- helping students to become autonomous learners
- reducing pressure on students by letting them work at their own speed
- enhancing students' literacy skills
- making teachers take a fresh look at the way they teach
- freeing teachers from administration to focus on students' learning

Although there is a significant literature on ICT in science education, much of it takes the form of articles on applications for use in teaching situations, with the emphasis more on how to use ICT, than on how to explore its effects. There is a sense in which it is taken rather for granted that ICT is a 'good thing': students are motivated when they use it, and this leads to better learning. Thus a central purpose of this review is to assess the strength of the evidence base to support the notion that the use of ICT activities in science lessons enhances students' understanding of science ideas.

1.2 Definitional and conceptual issues

ICT teaching activities have been taken to include tutorial applications, simulations, modelling, data-logging, graphing, use of multimedia, use of the internet. Activities which have been excluded are word-processing of essays and assessment record-keeping using ICT. The former has been excluded as essays are seldom used in science teaching, and the latter as it is a practical facility used by teachers not directly related to student learning.

Science has been taken to include one or several of the school science subjects: that is, integrated/ general science, science, biology, chemistry, physics and earth science.

Computer-assisted learning (CAL) and **computer-assisted instruction (CAI)** cover subject-specific software which provides students with instruction in the form of a tutorial-style programme on the material being covered.

Integrated learning systems (ILS) covers programmes which provide students with individualised instruction in the form of an interactive tutorial system.

1.3 Policy and practice background

The use of ICT in schools to support learning is pervasive. Such have been the possibilities envisioned for ICT that the 1980s saw money being invested in a number of countries on an unprecedented scale in initiatives aimed at getting computers into schools, a trend which continued into the 1990s.

In addition to policies focusing on getting computers into schools, there has also been legislation requiring teachers to use ICT in their teaching. For example, there is a requirement in the National Curriculum for England (DfEE/QCA, 1999) that ICT is incorporated in the teaching of all subjects. Partly as a result of this, significant funding has been made available to train teachers in the skills they need to make use of ICT in their schools and in their lessons. In England, for example, between 1992 and 2002, all primary and secondary teachers were required to undertake training under a New Opportunities Fund (NOF) initiative to improve their ICT competence.

Yet questions remain over the benefits. In England, Ofsted, the Office for Standards in Education (2001, 2002) has published reviews on the impact of government initiatives on standards and on literacy. Ofsted (2001, p 2) concluded that there is 'emerging evidence of a link between high standards across the curriculum and good ICT provision' but that the 'contribution of ICT to the raising of standards in individual subjects remains variable'.

1.4 Research background

Although there is a significant literature on ICT in science education, much of it takes the form of articles on applications for use in teaching situations; the emphasis has been on how to use ICT, rather than exploring its effects. The research literature is less extensive, and much of the research evidence which does exist is based on comparatively small-scale studies. Research which does exist can be grouped into three main areas:

- A number of studies have been carried out on *specific applications of ICT*, such as tutorial programmes, simulations and data-logging. Much of the early research into ICT was of this form, with the majority of studies tending to be small in scale. The focus of this work has been on students, with studies exploring effects on students' learning and development of skills. However, some additional evidence has also been gathered on aspects such as students' motivation, and the role of the teacher.

- Work has been undertaken on *more general aspects of the effects of ICT*. These studies have focused both on students (for example, in exploring possible links between the use of ICT and performance in national tests) and on teachers (for example, by looking in detail at the role of the teacher in lessons which make use of ICT).

- Studies have been undertaken into the *problems associated with the use of ICT* in science lessons and in schools more generally. Such studies have tended to focus on managerial and practical issues associated with the use of ICT.

Research into specific uses of ICT teaching activities in science lessons

One of the early uses of ICT was in providing CAL subject-specific software which enabled students to reinforce basic learning or, at a slightly more sophisticated level, provide students with instruction in the form of a tutorial-style programme on the material being covered. Development of these applications has resulted in ILS programmes which provide students with individualised instruction in the form of an intelligent tutorial system which provides almost immediate feedback on performance. Rogers and Newton (2001) explored its potential for supporting investigative work in practical science with 13- and 14-year-old students. Their work suggested that the software had been successful in promoting students' abilities to collect and manipulate data, but less successful at making links between the data they had collected and the associated science.

Several studies have explored effects of the use of simulations and modelling. Two studies with a focus on student learning were the *Conceptual Change* *in Science Project* undertaken in the UK (O'Shea et al., 1993; Hennesey et al., 1995), and the *Model-based Analysis and Reasoning in Science (MARS)* project, undertaken in the USA (Raghaven and Glaser, 1995). Both focused on aspects of students' misunderstanding of science concepts (a major area of research in science education) and both concluded that there was mixed evidence about effects on learning.

Data-logging (referred to as microcomputer-based labs, or MBL, in the USA) involves using electronic sensors during practical work to take measurements and then send them to a computer for processing. As data-logging and graphing were two of the earlier applications of ICT to be incorporated into science lessons, their effects have been researched in some detail. Earlier studies (for example, Nakhleh and Krajcik, 1993, in the USA) suggest gains in students' abilities to interpret graphs. More recent work in the UK by Barton (1997a, 1997b) does not appear to confirm this finding.

Advances in multimedia technology have resulted in CD-ROMs being developed for use in science lessons to allow students to perform 'virtual experiments'. Although some of these concentrate on practical activities which are difficult to do in the school laboratory, others have provided an alternative to normal practical work in science lessons. Collins et al. (1997) report on work done in exploring the effects of such software in a range of school subjects, including science, and report benefits in terms of understanding.

Research into more general effects of the use of ICT

A comparatively recent development in research has explored possible links between the use of ICT in schools and the standards achieved by students in national tests and examinations. Recent studies by the British Educational Communications and Technology Agency (Becta, 2001a, 2001b) have sought to compare the performance of students in schools well-resourced for ICT with those less well-resourced. The studies indicated that students at schools with 'good' ICT resources achieved significantly better results in national tests in English, Mathematics and Science at ages 11 and 14, and in national examinations at 16+ than students at schools with 'poor' ICT resources. Achievement was higher in schools where ICT was used routinely in mathematics and science lessons, and the best results were seen in schools where ICT was used across the whole curriculum.

Research into problems associated with the use of ICT

The scale of investment in ICT has been massive, and it is therefore scarcely surprising that a point was reached where people started to look at the 'returns'. Towards the end of the 1980s, questions were being asked from within and beyond the education sector about the ways in which ICT was being used in lessons, the expertise which had been acquired by teachers, and the knowledge and skills being acquired by students. Underpinning these questions was the concern that the impact of ICT has been less than had been anticipated, and the reality in the classroom was falling short of the aspirations of those promoting the use of ICT in schools. Evidence from surveys – such as those undertaken by McKinsey and Co. (1997), Goldstein (1997) and Poole (2000) – suggested that the problems in science lessons arose from a combination of educational and practical reasons. Although many of these were not unique to science, they were, arguably, brought more sharply into focus in science (and mathematics) lessons because these subject areas initially appeared the more natural 'home' for many ICT applications, and expectations were therefore higher. Reasons for problems being encountered included the following:

- doubts held by teachers over the value of ICT in promoting learning in science lessons
- the lack in many ICT resources of a clear rationale for their inclusion in teaching
- lack of adequate training for teachers
- a lack of time for teachers to plan for effective use of ICT in their lessons
- the planning difficulties associated with banks of networked computers being located centrally in rooms which had to be booked in advance
- teachers feeling threatened by the presence in the classroom of a new, powerful source of information
- lack of confidence on the part of many teachers with hardware and software
- shortage of computers
- lack of technical support
- unrealistic expectations about the nature and speed of change on the part of those implementing initiatives

More recent studies, such as those of Wellington (1999) and Newton (2000), through focusing on specific applications of ICT, have provided evidence of the persistence of many of the problems listed above. However, some evidence of improving teacher confidence has been provided by a survey in England and Wales undertaken to establish the impact of an investment of over £2.3 million in ICT training for teachers through the New Opportunities Fund (NOF) scheme. Data gathered showed that 73% of teachers reported themselves to be confident in the use of ICT in their teaching, compared with a figure of 63% two years previously (Department for Education and Skills (DfES), 2001), though these figures do not reveal anything of what happens in practice in lessons.

This brief overview of research into the use of ICT in science teaching supports the need for a systematic review into its effects on students.

1.5 Authors, funders and other users of the review

The Review Group all have an interest in both the substance of the review and the methodological approach of systematic reviewing. They are members of the EPPI-Centre Review Group for Science, and all have worked on several reviews for the EPPI-Centre and the TTA. Judith Bennett acts as a tutor on the Science Initial Teacher Training (ITT) programme at The University of York. A review of the use of ICT in science teaching forms one of the chapters in Judith Bennett's book *Teaching and Learning Science: A Guide to Recent Research and its Applications* (Bennett, 2003).

The project is funded by the TDA, which is concerned with bringing reviews of research literature to bear on the training of teachers. It is hoped that the results of this review will inform beginning and continuing teachers more fully about an important part of their subject. It is undertaken at this time as ICT has impacted considerably on the teaching and learning of school subjects in the last fifteen years. It is time to take stock of developments in the field.

The principal audiences for the review are likely to be teacher educators, trainee teachers and in-practice teachers. The review will also be of interest to teachers interested in research, policy-makers, researchers and students.

1.6 Review question

The main review question is as follows:

What is the effect of using ICT teaching activities in science lessons on students' understanding of science ideas?

The review focuses on students in the 11–16 age range, and on studies published in the period 2000–2004.

CHAPTER TWO

Methods used in the review

2.1 User involvement

2.1.1. Approach and rationale

Since this review is sponsored by the TDA, one of its main audiences will be teachers, trainee teachers and teacher educators. Accordingly, we used the PGCE Science cohort at the University of York in 2005/2006 for its development and dissemination. Specifically, a session was held in the autumn term 2005, when trainees returned from one of their teaching placements, to review progress on the review to date and to seek feedback. There are already projects underway in the department on the translation of research findings into teaching plans.

2.1.2 Methods used

In addition to the methods outlined below, the three review groups contracted to undertake reviews for the TDA (English, Science and Mathematics) hold joint meetings during the course of the three-year project to share good practice, both in the undertaking of the reviews and in their application with PGCE students. These review groups are constituted so as to reflect the range of potential users of review findings.

2.2 Identifying and describing studies

2.2.1 Defining relevant studies: inclusion and exclusion criteria

For a paper to be included in the systematic map, it had to report on a study looking at the effect of ICT teaching activities in science teaching on students' understanding of science ideas. As the focus of the study is on the effect of ICT, papers using methods to identify any such effects were required. Thus the review focused on evaluation studies: that is, type C studies in the EPPI-Centre taxonomy of study type contained in its core keywording strategy (EPPI-Centre, 2002a).

The review has been limited to the period 2000–2004 because of the rapid developments in the use of ICT in teaching.

Only studies published in English were included for pragmatic reasons related to available resources.

As the review included a number of science subjects in schools worldwide, it focused on students in the age bracket 11–16 (commonly the ages of compulsory secondary education), middle ability students and mainstream educational settings in order to contain the number of variables.

The full inclusion and exclusion criteria are contained in Appendix 2.1.

2.2.2 Identification of potential studies: search strategy

Reports were identified from the following sources.

Searching of electronic bibliographic databases

PsycINFO on 31 January 2005

BEI (British Education Index) via Dialog on 31 January 2005

SSCI (Social Sciences Citation Index) on 7 February 2005

ERIC (Educational Resources Information Center) on 9 February 2005

Keywords and descriptors for searching

Teaching, learning

Science education, science instruction

Science, physics, chemistry, biology

Computer uses in education, computer-assisted instruction

Information and communication technology, integrated learning systems

ICT, CAI, CAL, internet, multimedia, web-based
Secondary education

Higher degree theses were not included in the search as it is not possible to access these systematically; that is, it is not possible to access routinely theses in countries other than the UK. However, the search yielded papers arising from conference presentations on such work.

The full search strategy for the electronic databases is contained in Appendix 2.2.

2.2.3. Screening studies: applying inclusion and exclusion criteria

The Review Group set up a database system, using EndNote, for keeping track of and coding reports found during the review. Titles and abstracts were imported into the database. Inclusion and exclusion criteria were applied to (a) titles and abstracts, and (b) full reports. Full reports were obtained for those studies that appeared to meet the criteria or where there was insufficient information to be sure. The inclusion and exclusion criteria were re-applied to the full reports and those that did not meet these initial criteria were excluded.

2.2.4. Characterising included studies

The studies remaining after application of the inclusion/exclusion criteria were keyworded, using EPPI-Centre core keywording strategy: Data Collection for a Register of Educational Research, Version 0.9.7 (EPPI-Centre, 2002a). Additional keywords, which are specific to the present review, were added. All the keyworded studies were added to the EPPI-Centre database, REEL, for others to access via the website. Analysis of the keywords across all studies resulted in a systematic map of the review area, presented in frequency tables, a flowchart and cross-tabulations.

The generic EPPI-Centre keywords and the review-specific keywords are contained in Appendix 2.3.

2.2.5. Identifying and describing studies: quality-assurance process

Application of the inclusion and exclusion criteria and the keywording were conducted by pairs of the Review Group for a sample of studies, with each team member working first independently and then comparing their decisions and coming to a consensus.

2.3 In-depth review

2.3.1 Moving from mapping to in-depth review

Once studies had been keyworded and the systematic map generated, a meeting was held to reflect on the mapping of the field, and to decide whether any further inclusion/exclusion criteria might be applied. Such a narrowing down was a decision taken in the light of the overall review question and the aim of the review, as well as in terms of feasibility.

2.3.2 Detailed description of studies in the in-depth review

Following the creation of the map using the keywording data, it was possible to consider the in-depth review question. Inspection of the data allowed the development of a worthwhile and feasible in-depth research question based on good quality studies. The approach adopted was to consider the following:

- which type of ICT activity was mostly evaluated in the research reports (review-specific keyword 5)?
- which were the most appropriate study designs, that is, researcher-manipulated/controlled trials (generic keywords 13 and 14)?
- which studies involved pre-post achievement tests (review-specific keyword 8)?
- which studies involved representative or average/typical students (that is, not identified as gifted, less able or disaffected) (review-specific keyword 2)?

This gave an in-depth review question:

> What evidence is there from controlled trials of the effects of simulations on the understanding of science ideas demonstrated by students aged 11-16?

Studies were excluded from the in-depth review using the following criteria:

1. did not focus on the use of simulations
2. were not of a researcher-manipulated/controlled trial design
3. did not report pre-/post-test results
4. did not involve representative/average students

Studies identified as meeting the inclusion criteria, were analysed in depth, using the EPPI-Centre's detailed data-extraction review, Guidelines for Extracting Data and Quality Assessing Primary Studies in Educational Research, Version 0.9.7

(EPPI-Centre, 2002b) and online software, EPPI-Reviewer (EPPI-Centre, 2002c).

Data-extraction was completed online, using the EPPI-Centre guidelines and care was taken to answer each of the 93 questions for each study. The questions cover: aim(s) and rationale; research question(s), policy or practice focus; design; groups; sampling strategy; recruitment and consent; actual sample; data collection; data analysis; and results and conclusions, finishing with an overall assessment of the trustworthiness of the study to answer its own research question(s). The data-extraction of each study was completed independently by two reviewers. Then each pair of reviewers met to discuss, moderate and agree a final version.

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

Three components were identified to help in making explicit the process of apportioning different weights to the findings and conclusions of different studies. Such weights of evidence were based on the following:

A Soundness of studies (internal methodological coherence), based upon the study only

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

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

D An overall weight taking into account A, B and C

As part of the Science team's system for making consistent judgements on studies, a previously developed and successfully employed scoring system for each study was employed. This five-point scale allocated ratings from high (H), medium-high (MH), medium (M), medium-low (ML) to low (L) for five features relevant to weight of evidence (WoE) B and five features of WoE C. Along with the judgement made on WoE A (M11 from the data extraction process) these ratings contributed to an overall evaluation of each study (WoE D).

The weight of evidence indicators are contained in Appendix 2.5.

2.3.4 Synthesis of evidence

The data was synthesised to bring together the studies which answered the review questions and which met the quality criteria relating to appropriateness and methodology.

In order to carry out the synthesis, a summary report for each study (see Appendix 4.1) was drawn up, using key items within the EPPI-Reviewer data-extraction tool. These items were agreed among the core Review Group. Only one characteristic, details of the researchers, is not included in this tool but was considered important for the review; this information is included in the summary tables. These reports were edited by one team member for consistency of terminology, depth and detail, continuously referring to each relevant study. The reports were used by two team members to identify commonalities across the studies for the same characteristics as presented in the map. In addition, 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 for each study. 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.

Statistical synthesis was not carried out because the sample was heterogeneous in respect of science subject areas, country of research and hence educational system and variety of statistical tests adopted by the researchers.

The findings from the individual studies were clustered according to common features agreed by two members of the team. These are described and discussed in section 4.4.2 as the findings of this review (and summarised in 5.1.4).

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

2.3.5 In-depth review: quality-assurance process

Data-extraction and assessment of the weight of evidence brought by the study to address the review question was conducted by pairs of the Review Group, working first independently and then comparing their decisions and coming to a consensus. Members of the EPPI-Centre participated in data-extraction and weight of evidence assessment for a sample of studies as part of the quality-assurance process.

CHAPTER THREE

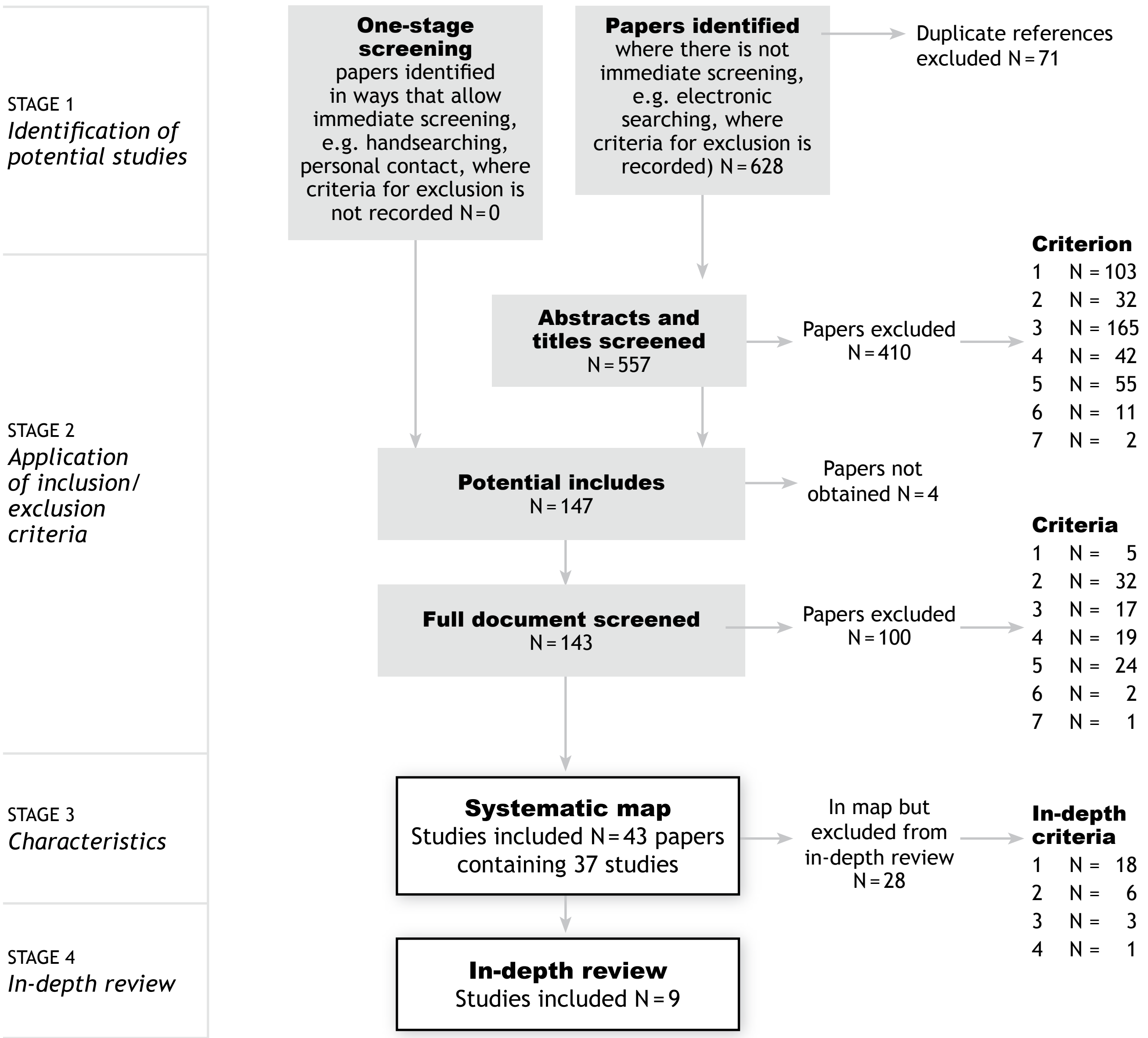
Identifying and describing studies: results

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 electronic searching described in section 2.2.2 yielded 628 papers of possible relevance. Within the 628 there were 71 duplicate references. The remaining 557 abstracts and titles were screened using the inclusion and exclusion criteria described in Appendix 2.1. Of these, 410 were excluded. The 147 potential includes were sent for and all but four (3%) of these were obtained for second screening. Of the full documents available for second screening, 100 papers were excluded by re-application of the inclusion and exclusion criteria. The remaining 43 papers contained 37 studies. Nine of the 37 studies met the criteria for the in-depth review described in section 2.3.2.

In the following tables, numbers and percentages will be given. In most cases, the numbers will total 37 as there were 37 studies included in the map (that is, the categories are mutually exclusive). However, in some studies more than one feature is reported on, for example, as shown in Table 3.4, students work in more than one way so the categories are not mutually exclusive. In all cases, the percentages are calculated on the basis of the 37 studies.

Figure 3.1 Filtering of papers from searching to map to synthesis



3.2 Characteristics of the included studies (systematic map)

As can be seen in Table 3.1, studies included in the map were carried out in ten countries. While the review focused on studies published in English, nearly half (46%) of them were carried out in countries where English is not the national/first/mother language. The remaining 20 studies (54%) were from the USA, the UK and Australia.

As was found in previous Science reviews, the highest proportion of studies was from the USA: 32% in Bennett et al. (2003) and 39% in Bennett et al. (2004).

The relatively high proportion of studies from Taiwan is in part due to the continuing work of one researcher who has published four studies (seven papers) on various pedagogical aspects of the ICT he has been developing (Chang, 2000, 2001a, 2003, 2004 as the main papers in this review).

One of the inclusion/exclusion criteria at the screening stage was that the studies should be evaluations; 27 (73%) in this review were found to be researcher-manipulated and ten (27%) naturally occurring evaluations. Within the researcher-manipulated group, eight (30%) were randomised controlled trials and 19 (70%) controlled trials. This balance reflects the limited opportunities within the educational setting for researcher(s) to have full control and to randomise which students or classes can be allocated to which treatments. In some cases, randomisation was limited to allocation of classes. However, where this occurred with less than four classes for each treatment the study was not classified as randomised for this review. This follows the recommendation of Ukoumunne et al. (1999) as to what constitutes a randomised trial or study when allocation is by class.

Table 3.2 shows that the biggest proportion of studies was Biology-based, with nearly as many in Physics. ICT seems to be little used, or little reported, in Chemistry teaching. This imbalance is consistent with an earlier EPPI-Centre systematic science review (Bennett et al., 2004) into the use of small-group discussions where only 4% of the studies involved Chemistry teaching and learning. On the other hand, in a systematic review focused on context-based and Science-Technology-Society approaches to teaching science, chemistry-based studies formed 23% of the sample (Bennett et al., 2003).

The close link between ICT and physics for the physics researchers, teachers and students can easily be appreciated. There are a number of reasons for the interest in ICT for biologists: the facility that it provides to avoid using animals for dissection (Akpan and Andre, 2000; Kariuki and Paulson, 2001); the manner in which it can speed up generation time when teaching evolution (Miglino et al., 2004); and the opportunity it gives to teach genetics in a more interactive way (Tsui and Treagust, 2003a).

As described above in relation to country, four of the seven earth science studies are those carried out by Chang (2000, 2001a, 2003, 2004) as his series of investigations into different aspects of the use of ICT in teaching students about the effects of typhoons and debris flow.

The most striking aspect of the data in Table 3.3 is that few authors give any details of the ability or motivational level of the learners taking part in their studies. As many educational systems are mixed ability, it is likely, unless specified otherwise, that the students would be of mixed ability for their age group or grade class. If this were the case, 87% of the studies reported on would be in this category.

Table 3.4 shows that 22 (60%) of the studies gave information on the ways in which students work with ICT, ranging from singly to whole classes. Within the sub-sample of 22 studies the most common practice (64%), was for single working followed by working in pairs (36%). It was unusual for students to work in larger groups.

A very high proportion of the studies was focused on measuring students' gains in scientific knowledge/explanations. Table 3.5 also demonstrates that nearly half of the investigations measured some aspect of students' learning about scientific approach / method. Twelve studies considered *both* scientific knowledge / explanations and scientific approach.

The cross-tabulations in Table 3.6 show that researchers in all five science subject areas were principally interested in the effect of ICT on scientific knowledge / understanding. The numbers by subject for scientific knowledge / explanations in Table 3.6 are very close to those for subject distribution in Table 3.2. It should be noted that the totals for scientific knowledge and scientific approach do not exactly match those in Table 3.5, 32 and 17 studies respectively, because one study (Dimitrov et al., 2002) involved Biology and Physics.

A considerably higher proportion of studies of the use of ICT in Integrated Science (3 out of 4) and Earth Science (6 out of 7) explore effects on understanding of the scientific approach than in Biology (5 out of 14) and Physics (3 out of 11), but the absolute numbers are very small for Integrated Science and for Earth Science, and so no conclusions can be drawn about the difference between the map sample and the in-depth sample for these two science areas.

How authors described their ICT activity varied somewhat. Nonetheless the frequency distributions in Table 3.7 show that simulations / virtual

Table 3.1
Country in which the study was carried out (mutually exclusive) (N = 37)

Country of origin	Number of studies	Percentage
USA	15	40
Taiwan	8	22
England/UK	3	8
Israel	3	8
Australia	2	5
Korea	2	5
Greece	1	3
Hong Kong/China	1	3
Italy	1	3
Turkey	1	3
Total	*37*	*100*

Table 3.2
Distribution by science subject areas (not mutually exclusive) (N=37)

*One study (Dimitrov et al., 2002) used a multimedia virtual environment to teach a biological topic and a physical topic.

Subject	Number of studies	Percentage
Integrated science	4	11
Biology	14	38
Chemistry	2	5
Physics	11	30
Earth science	7	19
Total	*38**	-

Table 3.3
Types of learners involved (mutually exclusive) (N=37)

Types of learners	Number of studies	Percentage
Mixed ability	7	20
Lower ability/slow learners	-	0
Upper ability/gifted	2	5
Disaffected	-	0
Unspecified	25	67
Other	3	8
Total	*37*	*100*

Table 3.4
How students work with the ICT (not mutually exclusive) (N=37)

Ways of working	Number of studies	Percentage
Singly in school	14	38
In pairs in school	8	22
In groups of up to 5 in school	4	11
Whole class together in school	1	3
Unspecified	15	41
Total	*42*	-

Table 3.5
Aspects of science understanding investigated (not mutually exclusive) (N=37)

Aspect of understanding of science	Number of studies	Percentage
Scientific knowledge/ explanations (facts, concepts, laws, theories)	32	86
Scientific approach (evidence, scientific methods, problem-solving)	17	46
Ideas about science (limitations, scientific community, risk, etc.)	1	3
Applications of science	1	3
Other	1	3
Total	*52*	-

Table 3.6
Relationship between area of science and aspect of science understanding studied (not mutually exclusive) (N=37)

	Scientific knowledge/ explanations	Scientific approach	Ideas about science	Applications of science	Other
Integrated Science	2	3	0	1	0
Biology	13	5	0	0	1
Chemistry	1	1	0	0	0
Physics	11	3	0	0	0
Earth Science	6	6	1	0	0
Total	*33*	*18*	*1*	*1*	*1*

Table 3.7
Types of ICT used in the studies (not mutually exclusive) (N=37)

Aspect of ICT	Number of studies	Percentage
Modelling	3	8
Simulations/virtual environments	19	51
Data-logging	4	11
Use of databases	4	11
Multimedia	10	27
Internet/WWW	7	19
Online discussion	1	3
Tutorial applications	6	16
Hypertext	2	6
Moving image (animations, video clips)	11	30
Email	2	6
Games/adventures	1	3
Other	7	19
Total	*77*	-

Table 3.8
Size of the study (mutually exclusive) (N=37)

Size of study	Number of studies	Percentage
Across several/many schools	4	11
One school, several classes	20	54
One class	4	11
Other	9	24
Total	37	100

Table 3.9
Methods used to measure outcomes (not mutually exclusive) (N=37)

Outcomes reported	Number of studies	Percentages
Pre- and post- test results	28	76
Test results	9	24
Examination results	3	8
Written reports /open questionnaires	18	49
Computer logs /computer products	5	14
Observed behaviour (including video)	8	22
Recorded group discussions (audio)	3	8
Interviews	9	24
(Dis)agreement scores	2	5
Total	85	-

Table 3.10
How were the studies published? (mutually exclusive) (N=37)

Type of paper	Number of studies	Percentage
Academic journal	29	78
Conference paper	7	19
Book chapter	1	3
Total	*37*	*100*

environments were the most common types of ICT used. In some cases, such as the studies of Chang (2000, 2001a, 2003, 2004), these simulations were provided together with other ICT facilities, including videorecordings and the internet. Thus the studies were also classed as 'multimedia' by the author. Where this was the case, the several terms were all keyworded to record the detail necessary for this review.

Science understanding was the only focus (dependent variable) for 14 studies (38%). Twenty-three (62%) also investigated other aspects, such as attitudes to computer-assisted learning; attitude to science topic; effort; cognitive preference; and implications for instructional practice.

ICT was the only independent variable investigated in 16 of the studies (43%) with 21 (57%) including other variables, such as gender; students' ability level; students' cognitive developmental level / stage; students' study strategy; authority level of the data presented to the students; students' spatial ability; students' prior knowledge/ understanding; students' motivational level; and feedback or not.

Table 3.8 shows that just over half the studies were carried out using several classes within one school. This has the advantage of controlling for some of the possible extraneous variables but rarely provides a substantial sample size.

Four studies were carried out on a larger scale and four utilised just one class.

The nine studies classified as 'Other' included ones where the number of classes was mentioned but no details given of schools involved or where student numbers were given but no details of classes and/ or schools.

The data in Table 3.9 shows that many studies utilised more than one method to measure science understanding. Three-quarters adopted the pre-/post-test approach, while written reports, which included questionnaires, were used in half the studies. Data was also gathered through test results (24%), observed behaviour including video (22%), and by interview (24%). Computers themselves were used in 14% of cases to log the activities of the students. Less frequent measures were examination results (8%), recording of group discussion (8%) and (dis)agreement scores (5%).

As can be seen in Table 3.10, the majority of the studies included in the map were published in academic journals.

The data in Table 3.11 demonstrates the relationship between the different types of ICT used and the types of science understanding being investigated in the 37 studies.

Table 3.11
Relationship between the ICT activity and the two principal aspects of science understanding investigated (N=17, not mutually exclusive)

Aspect of ICT	Scientific knowledge/ explanations (N = 32) %	Scientific approach (N = 17) %
CAI/CAL (no details on type)	0	0
Modelling	9	0
Simulations/virtual environments	53	53
Data-logging	6	24
Use of databases	9	24
Multimedia	31	29
Internet/WWW	16	29
Online discussion	3	0
Tutorial applications	9	24
Hypertext	6	6
Moving image (animations, video clips)	34	41
Email	3	12
Games/adventures	3	0
Other	19	18

Note: The relationship is expressed as a percentage of each separate aspect of science understanding. N refers to the number of studies that investigate each aspect of science understanding.

A number of observations can be made from Table 3.11.

The effectiveness of three types of ICT was tested in the same or similar relative proportions to teach both 'scientific knowledge/explanations' and 'scientific approach': simulations (53% and 53% respectively), multimedia (31% and 29%) and moving images (34% and 41%). However, as explained in relation to Table 3.7, in a few studies, the authors used the terms 'multimedia' as well as 'simulations' and/or 'moving images'. So, in the figures in Table 3.11, there is some overlap of these three ICT types and this will contribute somewhat to the similarity in their relative proportions.

Data-logging, databases, tutorials and email are more frequently evaluated for teaching to teach the 'scientific approach' rather than 'scientific knowledge / explanations'. This is not surprising as the use of data-logging and databases are scientific data-handling techniques, rather than the wider concepts of hypotheses, experimental design, and so on. Email was used in two studies to enable students to communicate with professional scientists and contribute data to wider meteorological projects.

Very few of these evaluation studies are focused on 'ideas about science', such as its limitations, the scientific community and risk.

3.3 Identifying and describing studies: quality-assurance results

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

Quality assurance of the two stages of screening papers retrieved from the electronic searches

First-stage screening

First-stage screening of titles and abstracts only was carried out in the Endnote database by three members of the Review Group. The 557 citations identified by the electronic searches were divided up roughly equally: 200, 200 and 157 between the team members, and annotated in accordance with the inclusion and exclusion criteria. Where there was any doubt about inclusion or exclusion, an inclusive approach was adopted.

As a check for reliability, the inter-screener agreement was assessed for two samples of 45 studies using the frequency and the Cohen's Kappa methods. The Cohen's Kappa method has the advantage of compensating for chance agreement. The results for the three screeners in two pairs are shown in Table 3.12.

Table 3.12
Inter-screener agreement (include-exclude, sample of 45)

	FREQUENCY METHOD		COHEN'S KAPPA METHOD	
	Identical decisions	**Inter-screener agreement**	**Cohen's Kappa coefficient**	**Inter-screener Agreement**
First screener: Second Screener	40	89%	0.638	Good
Second Screener: Third screener	44	98%	0.939	Very good

As a further quality-assurance check, the second sample was also screened by the EPPI-Centre link person. Inter-screener reliability was calculated by the frequency method only, but agreement with both screener 2 and screener 3 was very high, with identical decisions on 43 papers (95%) and 42 papers (93%) respectively. The differences in decision on the remaining papers in each case reflected a more inclusive approach by the EPPI-Centre link person.

Second-stage screening

Papers included at the first stage were then obtained and rescreened on the basis of the full paper by two members of the Review Group. The few discrepancies in the decisions of the screeners were discussed and agreed. A sample of five papers was also screened by one team member and the EPPI-Centre link person with 100% agreement.

Quality assurance of keywording

All 37 studies included in the map were keyworded by one member of the Review Group. Three different samples of five studies each were also keyworded by two other members of the Review Group and the EPPI-Centre link person for quality-assurance purposes, and then moderated.

Agreement was generally very high on all generic and specific keywords. Any discrepancies between decisions of the keyworders were discussed and resolved. Any papers that reported on the same studies were identified at this stage.

3.4 Summary of results of map

Thirty-seven studies met the inclusion criteria developed for the overall research review. These studies were keyworded and formed the basis of the systematic map. The map revealed a number of characteristics of research on the use of ICT in science education, as summarised below:

- The majority of the studies reported work that has taken place in the USA (40%) and Taiwan (22%).

- A little over one-third of the studies was on Biology topics and just under one-third on Physics topics. Very little research has been done in relation to Chemistry education.

- Few authors gave explicit details of the ability range of their participant students. (It was therefore assumed that students were mixed ability or average for their age unless otherwise stated.)

- In one-third of the studies, students worked individually with the ICT, and, in eight studies (22%), students worked in pairs. Two-fifths of authors (15 studies) did not give details of how the students interacted with the computers.

- Close to 90% of the studies focused on the students' understanding in respect of scientific knowledge/explanations and 50% on scientific approach (12 studies investigated both). This interest was spread across Earth Science, Biology and Physics.

- Types of ICT activities used varied, but half were referred to as simulations, either of experiments or of virtual environments. Virtual environments included a range of other ICT activities and non-ICT resources and could be defined as multimedia. There is therefore some overlap and flexibility in how the various forms of ICT activities are described or named.

- Half of the studies were carried out in one school with several classes. Only four studies (11%) involved large samples over several schools. Nine studies did not give full details of how many schools or classes were involved, although they all gave student numbers.

- Three-quarters of the studies used pre-/post-testing and half used questionnaires. Test results (that is, post- but no pre-test) were used in a quarter of the studies, as were interviews. Eight studies (22%) observed the student activities.

- Three-quarters of the studies were published in academic journals, seven (19%) as conference papers and one in a book chapter.
- Simulations/virtual environments, multimedia and moving images were used in the same or similar proportions to teach both scientific knowledge and scientific approach. (This is not too surprising, given the overlap in these three ICT categories). Data-logging, databases, the internet and tutorial applications were used more often to teach scientific approach than scientific knowledge.

CHAPTER FOUR

In-depth review: results

4.1 Selecting studies for the in-depth review

The application of the exclusion criteria specified in section 2.2.1 resulted in 37 studies (see section 6.1) which provided a substantial pool to consider for the in-depth review.

Five of these studies (Chang, 2001a; Chang, 2003; Dimitrov et al., 2002; Hsu et *al.*, 2001; Koroghlanian and Klein, 2004) were reported in linked pairs or triads of papers. One paper was selected as the lead paper for each study, but data in both or all three papers was drawn on for data-extraction purposes.

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

As described in section 2.3.2, care was taken to focus the in-depth review on those studies that

Table 4.1 Studies included in the in-depth review

Author	Date	Title
Akpan and Andre	(2000)	Using a computer simulation before dissection to help students learn anatomy
Chang	(2000)	Enhancing tenth graders' earth-science learning through computer-assisted instruction
Chang	(2001a)	A problem-solving based computer-assisted tutorial for the earth sciences
Chang	(2003)	Teaching earth sciences: should we implement teacher-directed or student-controlled CAI in the secondary classroom?
Diehl	(2000)	'Reasoner's Workbench' program supports students' individual and collaborative argumentation
Dimitrov et al.	(2002)	Changes in students' science ability produced by multimedia learning environments: application of the linear logistic model for change
Huffman et al.	(2003)	Using computers to create constructivist learning environments: impact on pedagogy and achievement
Huppert et al.	(2002)	Computer simulations in the high school: students' cognitive stages, science process skills and academic achievement in microbiology
Miglino et al.	(2004)	Using artificial life to teach evolutionary biology

Summary tables for each of the studies included in the in-depth review are contained in Appendix 4.1.

Table 4.2
Countries in which the nine studies selected for in-depth review were undertaken (N=9, mutually exclusive)

Country	Number of studies	Percentage	Study
USA	4	44.5	Akpan and Andre (2000) Diehl (2000) Dimitrov et al. (2002) Huffman et al. (2003)
Taiwan	3	33.5	Chang (2000) Chang (2001a) Chang (2003)
Israel	1	11	Huppert et al. (2002)
Italy	1	11	Miglino et al. (2004)
Total	*9*	*100*	

Table 4.3
Number and percentage of studies in different science areas (N=9, not mutually exclusive)

Subject area	Number of studies	Percentage	Study
Integrated Science	1	11	Diehl (2000)
Biology	4	44	Akpan and Andre (2000) Dimitrov (2002) Huppert et al. (2002) Miglino et al. (2004)
Chemistry	-	-	None
Physics	2	22	Dimitrov et al. (2002) Huffman et al. (2003)
Earth Science	3	33	Chang (2000) Chang (2001a) Chang (2003)

Note: Dimitrov et al. (2002) looked at a Biology and a Physics unit.

most closely fulfilled the objectives of the review question:

- Which aspect of ICT was of most interest to researchers?
- Which were the most appropriate in design for an evaluation (researcher manipulated/controlled)?
- Which studies involved a pre-/post-design?
- Which studies involved representative or average/typical students?

The aspect of ICT of most interest to researchers proved to be simulations (19 of the 37 studies). This gave an in-depth review question:

What evidence is there from controlled trials of the effects of simulations on the understanding of science ideas demonstrated by students aged 11-16?

Once the other criteria were applied to those studies that focused on simulations the in-depth sample was reduced to nine studies (Table 4.1).

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 in-depth review (country of study and science subject focus) 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 study

Table 4.2 shows the countries in which the nine studies selected for in-depth review were carried out.

The proportion of those carried out in the USA was very close to that in the systematic map (40%). However, the proportion in Taiwan (33%) was greater compared to the map (22%), as was that of Italy (1% in the map sample). Given the small sample size for Italy and Taiwan, these differences are not unexpected.

Subject focus

Table 4.3 shows the distribution of the studies in relation to the science subject area.

In three subject areas, there was little or no difference between the proportions in the map and in the in-depth sample: Integrated Science was exactly the same (11%); Biology was 38% in the map; and there were no studies using Chemistry, which was close to the low proportion (5%) in the map. There was a somewhat lower percentage for Physics (22% in the in-depth sample, compared with 30% in the map). The greatest difference was with the studies in Earth Science (33%) compared with 19% in the map.

Overall, this section indicates that the nine studies in the in-depth review are representative of those in the systematic map in terms of reflecting country of study and science subject focus. Little comparison can be made in relation to ability level as no information is given by 68% of the map sample and similarly (78%) of the in-depth sample.

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

4.3.1 Overview of studies

This section describes various aspects of the nine studies and looks for patterns. It finishes with details of the assessment of weights of evidence and demonstrates how seven of the nine studies were considered of sufficiently high standard to contribute to the findings of the in-depth review.

Researchers – level of experience and independence as evaluators

As it is rare for authors to declare their status (for example, doctoral student, teacher-researcher, university-based academic), it can be difficult to assess their level of experience and the relative contribution of each author/researcher. The absence of such explicit information means that an important part of the context in which the study is undertaken is omitted. It is also not always clear how independent the researchers are in relation to the intervention they are evaluating. However, certain clues are sometimes available to assist with these assessments and the following information on the researchers has been deduced.

Of the nine studies, all but one appeared to have been undertaken by established university staff members. The exception was Diehl (2000), where the research was part of a dissertation for a US university and was, therefore, post-graduate work. Diehl was an independent evaluator of the 'Convince Me' programme which was developed by others in 1994.

The Akpan and Andre (2000) study was carried out by members of universities in two different states in the USA. Andre had published in the topic area since 1995 and Akpan had recently conducted a review of the literature which was published subsequently (Akpan 2001). The researchers were evaluating a commercially available dissection simulation programme, *BioLab Frog*.
Chang, at a university in Taiwan, worked singly on a series of studies (2000, 2001a, 2003) using a multimedia programme, including simulation, to investigate the effectiveness of various pedagogical approaches. The reference list showed that Chang had published on the topic since 1999. In these studies, he was evaluating the multimedia approach that he had designed and refined over several years.

Dimitrov, from one US university, worked with McGee and Howard at a second US university where the ICT programme had been developed. McGee and Howard had published on the topic of secondary school science education in 1999.

Huffman, Goldberg and Michlin collaborated across US universities in three different states. Goldberg had previously published (1995) on the design and development of the ICT simulation being evaluated.

Huppert and Lazarowitz worked respectively at a university and at a technicon in Israel in collaboration with Lomask at a US university. Hupper and Lazarowitz had published at least three joint papers on computer-assisted learning since 1986. Their study evaluated a software programme developed by Huppert and Lazarowitz, based on other researchers' earlier work.

Miglino, Rubinacci and Pagliarini, from an Italian university, worked with Lund at a Danish university on their pilot study of artificial life software for teaching evolutionary biology. All four had previously published in the topic area from 1996. Of the three software programmes investigated, two had been developed by one or more of the authors and the third was a commercial product published in Italy and widely distributed in schools.

Focus of studies ICT activities employed

Akpan and Andre (2000) In the context of the ethical concerns in the USA about animal dissections in classrooms, this study examined the use of simulation of frog dissection in improving students' learning of frog anatomy and organ function. The seventh-grade students worked singly or in pairs on the *BioLab Frog* software to carry out frog dissections.

The software is a simulation of a frog dissection supplied by Pierian Spring. It simulates on the screen an actual frog dissection. As the student views and removes organs, the software display adds information about each item. It incorporates *QuickTime* movies and microscopic pictures to illustrate functions that are normally hidden from view.

Chang The three Chang studies used multimedia computer facilities, which include guided inquiry, animated weather-satellite images, virtual field trips and internet usage. These activities were employed to evaluate different pedagogical approaches.

Chang (2000) The author investigated the comparative efficiency of computer-assisted instruction (CAI) and traditional teaching methods in Earth Science classes in Taiwan. The focus of learning was on knowledge (the recall or recognition of ideas or concepts), comprehension, and the students' ability to apply acquired knowledge to a new situation. Guided enquiry provided by a computer programme allowed tenth-grade students to work individually with a range of provided resources, video, animated weather maps, books, and so on in a virtual research office to prepare a research report on debris flow hazards following a typhoon.

Chang (2001a) This study was a development of earlier work (see Chang, 2000, above) and was formalised as a problem-solving computer-assisted tutorial. The subject matter was the same, but the control was different and involved lecture-internet-discussion teaching. The focus of learning was similarly on the recall or recognition of ideas or concepts, comprehension and the students' ability to apply acquired knowledge to a new situation. The software, which included relevant data, a virtual field trip and animated weather maps, provided guidance for interactive investigation. Students in the comparison group were given clear and detailed instruction and explanations by the teacher on the same topic, and used the internet to control for 'computer-novelty effects'.

Chang (2003) Building on previous studies Chang compared the achievements of tenth-grade Taiwanese students who experienced teacher-directed CAI (TDCAI) with those who undertook student-directed CAI (SCCAI). Both groups used the multimedia CAI software, which was designed to allow users to navigate the various learning sections in a non-linear fashion. The TDCAI approach emphasised direct guidance from the teacher, while the SCCAI stressed student self-paced learning.

Diehl (2000) This study evaluated the effectiveness of computer-mediated support for students' individual and collaborative argumentation with ninth-grade students in integrated science classes in the USA. *Convince Me* aids students in generating and analysing arguments, and provides feedback on argument coherence. Students can work individually to build arguments and thus obtain benefits that are often associated with collaborative activity. In this study, students worked individually or in pairs, with or without feedback.

The *Convince Me* simulation interface structures an argument by breaking down the process of building an argument into steps that identify hypotheses and evidence, as well as the explanatory and contradictory relations that join them. The software can be used with or without the feedback mode.

Dimitrov et al. (2002) The researchers investigated the potential of the *Astronomy Village: Investigating the Solar System* programme, which provides virtual mentor guides to support students in completing multiple investigation cycles that mirror the phases of scientific inquiry. The evaluation compared alternative uses of technology for teaching the same interdisciplinary content. The comparison focused on access to image analysis activities versus no access to image analysis activities. A no-treatment group was also used to allow for any changes due to maturation and experience. The evaluation was carried out on students at different grade levels across the USA.

The ICT activities in the virtual village based in Hawaii include a virtual mentor, datasets, hands-on activities, lectures and library articles. After completing their investigations, students host a press conference in front of virtual press corps.

Huffman et al. (2003) This study considers the extent to which computers can be used to create a constructivist learning environment in the science classroom. The evaluation focused on Motion and Force units within a larger *Constructing Physics Understanding* (CPU) project, which uses computer-based modular curricular activities, software and pedagogy to teach a number of physics topics. Students in 23 high-school classes in the USA were involved and classes were taught by teachers with one of two levels of experience with the software (experienced or beginner) or by non-CPU teachers.

The Motion and Force software package provides a set of computer activities, computer simulations and an electronic journal. Students develop, test

Table 4.4
The primary aims/pedagogic foci of the nine studies including the controls employed

Pedagogic focus	Comparisons/components	Study
Mainly on the ICT	Relative effectiveness of computer simulation and experience of hands-on frog dissection (traditional teaching)	Akpan and Andre (2000)
	Comparative effectiveness of computer-assisted instruction and traditional teaching on typhoon damage	Chang (2000)
	Relative effectiveness of alternative uses of technology (Astronomy Village) to teach the same interdisciplinary content.	Dimitrov et al. (2002)
	Computer-assisted learning of the life cycle of micro-organisms compared with traditional teaching	Huppert et al. (2002)
	Two artificial life programmes that give students experience in virtual experiments compared with use of hypertext only	Miglino et al. 2004
On the ICT and the instructional approach	Problem-solving based computer-assisted tutorial compared with lecture-internet-discussion in learning about typhoon damage	Chang (2001a)
	Student-controlled CAI compared with teacher-controlled CAI in learning about typhoon damage	Chang (2003)
	Use of software that provides feedback in argumentation (and thus can give individual students opportunities for collaborative working) compared with software but no feedback	Diehl (2000)
	Constructivist learning/level of teacher experience comparing three levels of teacher experience (experienced, beginner and no ICT/traditional) in using computers to teach physics	Huffman et al. (2003)

and modify their ideas through experimentation.

Huppert et al. (2002) A software programme was evaluated for its potential to enhance tenth-grade students' understanding of the life processes of micro-organisms. Control students studied the same learning material in the classroom and the laboratory. The study was carried out with tenth-grade biology students in Israel and the researchers took into account their cognitive stage (type of reasoning adopted by the student) in analysing their data.

The *Growth Curve of Micro-Organisms* simulation programme makes it possible to perform 'experiments' in short time and to check the influence of various factors, such as the initial number of organisms in a population, the temperature range and the nutrient concentration on the growth curve. It also gave opportunities to practise and improve science process skills. This was integrated into the sequence of learning activities in the classroom and laboratory, and was performed at the student's own pace.

Miglino et al. (2004) The researchers carried out a pilot evaluation on the use of artificial life software with high-school students in evolutionary biology classes in Italy. Two ICT educational packages were tested and a commercial hypertext product was used for the control students. All students received two standard lessons in evolutionary biology before using the software. A multiple-choice questionnaire was then used to test student knowledge. The focus of the evaluation was to test the importance of information with structured guidance rather than unstructured information.

One of the artificial life simulation tools was *Face-It,* a user-guided algorithm that pilots the evolution of facial expressions and was used to teach the concepts of artificial selection, genotypes and mutations. The other was *ToyBot*, which allows users to simulate interesting mobile robot behaviours. This model simulates the evolution of populations of artificial agents with elementary navigational possibilities and was used as a teaching aid to focus students' attention on the role of environment and individual behaviour in determining behaviour (the phenotype).

The control group of students used a commercially available hypertext product, *Charles Darwin,* with seven sections which include detailed text descriptions, maps, pictures and animations to describe basic concepts in Darwinian evolution.

Aims and the nature of the comparisons/ controls

In some cases, the researchers were using the ICT as a tool to investigate pedagogical issues as well as the specific contribution of the software. An example of this is the work of Chang, who used essentially the same multimedia resources to investigate questions relating firstly to guided inquiry via a computer program by way of contrast to traditional teaching (2000), then a problem-solving approach (2001a) and then student self-paced, independent working (2003). In this series of studies, he appears to be seeking the most effective way of using of the multimedia facilities. However, in a number of studies, it is difficult to separate the contribution of the ICT and the contribution of the instructional approach. Table 4.4 summarises the primary pedagogic foci of the nine studies. It should be noted, however, that these distinctions are not entirely sharp, but rather a question of relative balance.

4.3.2 Methodological considerations (weight of evidence A)

Weight of evidence A was based on the outcome of the data-extraction process and specifically M11 '*Taking account of all quality assessment issues, can the study findings be trusted in answering the study question(s)?*' The following section is a summary of the features of each study that were considered important in making the judgements.

Sample size and sampling method

Sample sizes were reasonable in most studies and these were determined in part by recruitment strategies and the nature of the ICT being evaluated.

The study with the largest sample (Dimitrov et al., 2002) with 837 students, (numbers of schools, classes and teachers not given) invited schools across the USA to participate in the evaluation and then selected their sample on given criteria. Similarly the Huffman et al. (2003) evaluation involved a cross-USA project with 23 schools, 13 teachers and 366 students. In this study, schools and teachers already involved with the Constructing Physics Understanding project were recruited to be involved in the evaluation and a rough balance of teachers for three levels of experience with the software was selected.

The three studies of Chang (2000, 2001a and 2003) utilised 151, 137 and 232 students respectively. In each case, there was one teacher and one school. It is not stated whether it was the same school and teacher over the four years, or how the school(s) or students were recruited. The 2003 study did include random allocation of six classes into each of the experimental and control groups.

The evaluation carried out by Huppert et al. (2002) involved 181 students from five classes, with no detail about school or teacher numbers.

Diehl's 2000 evaluation started with 127 students from across four ninth-grade classes and two teachers (no details about school numbers). However, students who did not attend a required minimum number of the sessions were excluded from the data set and the analysis was carried out on the data from 102 students.

The two studies with the smallest samples were those of Akpan and Andre (2000), and Miglino et al. (2004). The former started with 127 eligible students but for a number of reasons (illness, lack of baseline ability data, and so on) was reduced to 81 students with one teacher in one school. Miglino et al. (2004) gave little detail of their sample except that there were 22 students in each of the experimental and control groups and one teacher.

None of the studies in the in-depth review used an explicit sampling frame, such as a roll of students in the school, the list of classes in a school or a national or regional register of schools. However, as described above, a number were careful in choosing and/or balancing their samples (usually by class or teacher) and two did not use data from students who had not participated fully for whatever reason.

Methods used to collect the data and justifications for their employment

There are three main possible options for collecting data to assess students' understanding of science pre- and post-intervention. These cross the spectrum from employing already established and thus previously validated tests, or parts of such tests, via modifying established tests to fit the particular evaluation, to developing unique assessments for the specific intervention. All three examples were found in this review; in most cases, just one type per study. The Huppert et al. (2002) evaluation used two of these methods for different aspects of their study. In most studies, reasons were given for the choice.

Use of established tests

Huffman et al. (2003) used the nationally (USA) recognised and published *Force Concept Inventory* to measure students' understanding of force and motion in the *Constructing Physics Understanding* evaluation. The authors explain that they used this because it includes a wide range of concepts taught in a typical Physics class, it has good reliability, and because there are existing national data on how students in other high school Physics classes have scored on the test.

In order to assess student understanding of the scientific approach (such as measurement, interpreting data, prediction, and so on), Huppert et al. (2002) employed the previously published *Biology Test of Science Processes*. They used this test because five high-school teachers found that it suited tenth-graders and represented the science process skills practised by the students with the CAL software.

Use of established test(s) modified for the specific evaluation

In assessing the value of the *Convince Me* software, Diehl (2000) adapted measures from three established tests. The one most relevant to this review was the *ThinkersTool Epistemology Assessment*. (The other two tests related to beliefs about science and perceived relevance of science.) However, for the evaluation, questions about science reasoning (defining evidence, hypotheses and defining scientific reasoning) were used together with questions about science to real world applications (belief and attitude statements), and these were analysed together. The post-tests also included an argumentation knowledge test. No details were given of its origin, although a sample of the questions was given in an appendix. No reasons were given for the choice of tests.

Use of tests designed specifically for the evaluations

Akpan and Andre (2000) developed a 25-item multiple-choice test and short answer instrument for their evaluation, using the expertise of one biology teacher and two 'science experts' with teaching experience. No reason is given for this approach and it is assumed that the test was focused on the specific material. (The *BioLab Frog* dissection software includes a short review quiz that matches each function to structure after each system. This was not used for assessment.)

In his three studies, Chang (2000, 2001a, 2003) used an Earth Science achievement test (ESAT) that he and a high-school teacher first developed for the 2000 study. This was a 30-item multiple-choice test with questions grouped into three cognitive levels: knowledge, comprehension and ability of students to transfer acquired knowledge to a new situation. While no explicit reason was given for this approach, it is implicit that the test was designed to suit the teaching material.

Dimitrov et al. (2002) were of the view that 'Very few large scale items focus on the kinds of inquiry-related performance expectations promoted by Astronomy Village', so they developed their own assessment instrument 'attuned to the topics and performance expectation in Astronomy Village' (p 17).

The study by Huppert et al. (2002) on students' understanding about the growth of micro-organisms does not describe the origin of their science knowledge test but does explain that the validity of the different pre- and post-tests was checked. From this, the reader could assume that the test was developed for the evaluation.

In assessing understanding of evolutionary biology, Miglino et al. (2004) developed their own 14-question multiple-choice test. They deliberately chose to use a simple tool because it would take up little of the students' time and would 'avoid psychological stress due to memory and attention overload' (p 124).

Reliability/validity/trustworthiness of the data-collection tools

The approach to establishing validity and reliability depended in part on the source of the tests themselves. Thus Huffman et al. (2003) and Diehl (2000) relied on using a well established and already validated, reliable test and did not do their own checks. Huppert *et al.* (2002) were extra cautious in taking an established test and checking it for validity and reliability for their study.

Researchers who developed their own tests needed to make their own checks for validity and reliability. As most of the pre- and post-intervention tests assessed gain in science knowledge and/or science process skills that were largely quantifiable, most authors were able to apply standard approaches for reliability and validity.

A number of studies used a panel of experts to check the content validity of their test questions and statistical checks for inter-rater reliability for test marking. For instance, for the simulation frog dissection study of Akpan and Andre (2000), the pre-/post-multiple choice test questions were designed by three scientists experienced in teaching. (Dissection performance was also assessed but not reported on.) The authors used the Cronbach's alpha test for internal consistency / reliability.

Also, Chang (2000, 2001a, 2003) used a panel of specialists (three high-school teachers and three professionals) to validate the test questions he had developed along with a high-school teacher. He then used the Kuder-Richardson Formula 21 to determine the reliability coefficient (0.77). To apportion the questions to the three different

Table 4.5
Weight of evidence A (WoE A) (N=9 studies)

	Study	WoE A
1	Akpan and Andre (2000)	ML
2	Chang (2000)	M
3	Chang (2001a)	MH
4	Chang (2003)	MH
5	Diehl (2000)	L
6	Dimitrov et al. (2002)	MH
7	Huffman et al. (2003)	MH
8	Huppert et al. (2002)	M
9	Miglino et al. (2004)	ML

types of science understanding (knowledge, comprehension and application of acquired ideas to new situations), the author used a panel of judges who were 'knowledgeable about the criteria of these categories' of science understanding. The reliability was then checked with a Pearson product moment coefficient to find a high level of agreement.

Dimitrov et al. (2002) employed item (question) writers to develop multiple-choice assessment items related to the underlying concepts within the intervention investigations. No mention is made of tests for validity. Cronbach's reliability check was used for content understanding and problem-solving items.

Miglino et al. (2004) developed the multiple-choice questions along with Biology teachers but no mention is made of tests for validity. The simple multiple-choice questions would not present a problem for the reliability of the marking, but these tests were not carried out.

Methods used to analyse the data and justifications

In all studies, standard statistical tests were employed. For the most part, these were analyses of variance that compared pre- with post-data for interventions and controls: Akpan and Andre (2000); Chang (2000, 2001a, 2003); Diehl (2000), although the author referred to the test as a 'regression'; Huffman et al. (2003); Huppert et al. (2002); Miglino et al.(2004). As these are the classic statistical analyses for this situation, most authors did not explicitly justify their validity.

Dimitrov et al. (2002) did not use ANOVA but rather the Linear Logistic Model for Change as this 'eliminates drawbacks of the traditional pre-test - post-test design, provides information about the magnitude of the change on a ratio scale, and separates changes due to treatment from changes due to natural trends across time points of measurement' (p 18).

In several studies, some additional tests were employed. Effect size was employed by Chang (2003) and Huffman et al. (2003). Chang (2001a) also tested for normal distribution with a Kolmogorov-Smirnoff test and used Bartlett's chi-squared test for homogeneity of variance among groups. Huffman et al. (2003) used a Hake Plot, which is a plot of percentage gain, as this can take into account that students with lower pre-test scores have more opportunity to gain than students with higher pre-test scores. Huppert et al. (2002) compared means scores of students with three cognitive levels pair-wise with a Sheffe test.

Reliability and validity of the data analyses

The standard statistical tests described above are valid and reliable for the studies. In addition, Diehl (2002) ensured that the free answers for the argumentation test were coded by two trained researchers and agreement for conflicting categorisation (less that 5%) was negotiated.

Summary of weight of evidence A (WoE A) judgements

The summary of the agreed weights of evidence for A for the nine studies is given in Table 4.5. The rating system is on a five-point scale as described in section 2.3.2: high (H), medium-high (MH), medium (M), medium-low (ML) and low (L). The same five-point scale was used for weights of evidence B and C.

4.3.3 Appropriateness of studies' research design for in-depth review (WoE B)

The following five aspects of each study were examined to reach a view on the appropriateness of the study design for the in-depth review question:

Sample size and sampling method

Sample size in most of the studies was sufficient or better for the purposes of this review. Three classes or about 90 students or fewer was seen as a small sample. Only the Miglino et al. (2004) study had starting numbers below this, with very limited information about those students other than their age. In two studies, Diehl (2000) and Akpan and Andre (2004), retrospective adjustments were made by excluding data from students who had not experienced a sufficient proportion of the intervention to justify inclusion. In the Akpan and Andre (2004) study, this brought the numbers down to 81. The Diehl (2000) study retained 102 students after losses for insufficient attendance.

In all cases, the participants in the studies were chosen by classes, teacher or school, rather than as individuals by means of a sampling frame. This is commonly the case for research in school settings when classes are chosen for opportunistic reasons and the researchers are rarely in the position of determining which individual students participate. Thus it was anticipated that sampling frames would not feature but should not be excluded as possibilities. The Dimitrov et al. (2002) study took care in selecting from volunteering teachers to ensure a demographic balance of students. This is one way of balancing extraneous variables (see below).

Selection of control

Once again the constraints on educational research mean that random allocation of individual students to experimental and control categories can rarely be achieved.

No study in this review reported random allocation of students.

Some researchers were able to allocate classes randomly to the two conditions. Akpan and Andre (2000) randomly assigned 'class periods' to four conditions, but gave no indication of how many classes that involved (starting student numbers were 127). Chang (2000, 2001a) gave no explicit details of allocation, but reported that he used the method of Campbell and Stanley (1966), which requires random allocation to control and experimental groups. However it is not clear from these two studies if this was applied to individuals or classes. The Chang (2003) study describes the random allocation of six classes to two conditions. However, allocation of fewer than four classes per condition is not considered to be fully random in the health field (Ukoumunne et al., 1999). Thus these studies may be considered quasi-random.

Control of extraneous variables

Full control of extraneous variables is not always possible in opportunistic or semi-opportunistic education research in which researchers select teachers/schools from a pool of those volunteering to be involved in the evaluation. However, many of the studies in this in-depth review demonstrate how some control may be achieved.

Teacher/school
One example is the way in which researchers control for teacher and/or school variability. Chang (2000, 2001a, 2003) approached this by using one teacher in one school for the experimental and control conditions. This approach is common and perhaps inevitable where small to medium numbers of students or classes are involved (Akpan and Andre, 2000; Miglino et al., 2004). In another study (Diehl, 2000) involving two teachers and four classes (presumably in one school), each teacher taught half the students in each of the four comparison groups. A comparison non-ICT control group was also used, but no details are given of the teacher.

In the larger cross-school studies, controlling for teacher and school effect is more difficult but not impossible. This can be achieved by matching teacher experience (global experience perhaps in years of teaching) or specific experience of the particular ICT. The CPU Project (Huffman et al., 2003) set out specifically to look at the importance of teachers' experience of the software by using 'lead' experience teachers, 'beginning' CPU teachers and non-CPU teachers covering the same topic material. In this study, the non-CPU teachers within the same schools were matched to the 'beginning' CPU teachers according to the teaching experience and the demographic characteristics of their students.

The other large cross-school evaluation study, Dimitrov et al. (2002), selected teachers for the two treatment conditions on the grounds of minimum hardware and software requirements, and the demographic characteristics of their students (as described below). However, each participating teacher was asked to recruit another teacher at their school for the 'no treatment' control group. Thus the school effect was taken into account.

The Huppert et al. (2002) study used three teachers to teach the two experimental and three control groups, but no further details are given and it is not clear whether the evaluation was carried out in one school.

Gender
When using classes for education research, it is not possible to control for gender by selecting equal numbers. Nonetheless, the contribution of gender may be derived by means of the statistical analysis. This was done in one of these studies, Huppert et al. (2002), where there was a preponderance of girls in the sample.

Akpan and Andre (2000), Chang (2001a, 2003) and Dimitrov et al. (2002) gave details of the gender balance but did not include it as part of their analysis. The Chang (2000), Diehl (2000), Huffman et al. (2003), and Miglino et al. (2004) studies gave no details on the gender balance of their participants.

Table 4.6
Weights of evidence B (WoE B) (N=9 studies)

	Study	WoE B
1	Akpan and Andre (2000)	MH
2	Chang (2000)	M
3	Chang (2001a)	MH
4	Chang (2003)	MH
5	Diehl (2000)	M
6	Dimitrov et al. (2002)	M
7	Huffman et al. (2003)	MH
8	Huppert et al. (2002)	M
9	Miglino et al. (2004)	ML

Demography
Demographic characteristics will only be an issue with large-scale, cross-school studies. Huffman et al. (2003) matched non-CPU teachers to 'beginning' CPU teachers for demographic characteristics and also ran statistical tests to check for such differences and found none. They gave no details of the characteristics they were measuring. Dimitrov et al. (2002) gave details of the ethnic composition of their sample and reported that all three conditions were balanced for this variable.

Ability
In order to produce findings that would be representative of the majority of students and from which generalisation might be drawn, students described as gifted or less able were excluded from the in-depth sample. However, several studies give no detail about the ability range or achievement level of the students (Chang, 2003; Diehl, 2000; Huppert et al., 2002; Miglino et al., 2004). The assumption has been made that it was implicit in these studies that the students were of mixed or average ability. This was because many classes are of this type and it was expected that researchers would be explicit if they were working with the extremes of the ability range. For example, an excluded study (Shim et al., 2003) worked with gifted and talented students in the top 1% of the ability range.

Huffman et al. (2003) checked for differences in student characteristics but did not make it clear what these characteristics were. Other studies make limited reference to balancing student ability range across treatment groups. Akpan and Andre (2000) had been able to roughly equalize ability [of students] across sections [class periods] at the beginning of the academic year and so for the four conditions. The Chang (2000) and (2001a) studies describe the participants as being 'typical of tenth grader students'. Dimitrov et al. (2002) list the grade level breakdown of the students and explain that the grade levels were proportionally balanced for the three treatment groups.

However, Huppert et al. (2002) did take certain aspects of ability into account in the analysis of their results. They were interested in students' cognitive stages and measured these at the start of their study, dividing the students into 'Concrete', 'Transitional' and 'Formal' reasoners (concrete being the least sophisticated reasoners). They then analysed the experimental and control data taking these stages into account.

The trustworthiness of the data collection

Almost all the studies were characterised by high levels of trustworthiness in relation to the methods of collecting data. This was achieved by either using established and published tests, by adapting appropriate parts of such tests or by developing in-study checks for reliability and validity. Only the Miglino et al. (2004) study did not employ any system for checking the reliability of their methods of data collection.

The trustworthiness of the data analysis

All studies used standard and appropriate tests to analyse their data. Usually these were ANOVA tests and some researchers utilised further statistical tests to refine their data analysis as described for weight of evidence A.

Applying the above five criteria to each study produced the weights of evidence B (WoE B) displayed in Table 4.7.

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

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

- the ways in which the students worked with the simulations
- the focus of the intervention
- the measures employed to test the nature of science understanding
- the breadth of the science understanding reported

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

The ways in which the students worked with the simulations

This considered the learning experience for the students on the basis that students would gain most from working interactively with the computer and in small groups, usually pairs. This allows for peer support (Vygotsky, 1962, 1978). Working alone, but interactively, was seen as a somewhat less valuable experience, while viewing simulations as a passive audience while another (that is, the teacher) demonstrated the simulation would be the least productive learning experience.

Not all the study reports gave details of how the students worked with the simulations. The Dimitrov et al. (2002) study was large-scale research across US schools so the way 720 treatment students worked with the Astronomy Village may have varied from school to school. The Huppert et al. (2002) and Miglino et al. (2004) studies were smaller scale, but did not include details of individual or group working arrangements.

In the Akpan and Andre (2000) study, the students worked individually with the frog dissection simulations (although in pairs when carrying out the real life dissections). Similarly in the Chang (2000, 2001a) studies, all students worked individually in a virtual private research office that includes simulations, such as animated weather-satellite images and virtual field trips. However, the third Chang study (2003) focused on the contrast between teacher-directed and student-directed use of multimedia resources (including simulation). In this comparison, the teacher-directed students experienced the simulation as a whole class audience (interactively but with the teacher manipulating the simulations), while the student-directed group worked individually, interacted directly with the software and was self-paced.

The Diehl (2000) study sought to evaluate the effectiveness of computer-mediated support for students' individual and collaborative argumentation about scientific process. A feedback facility in the software simulates collaborative activity and so provides a 'computer partner' in place of a 'student partner'. The four treatments involved students working individually or in pairs, with or without feedback with Convince Me.

Students worked in 'small groups' at computer stations for the CPU project (Huffman et al., 2003) but little other detail is given. Possibly group size varied from school to school across the sample of 23 US schools.

The focus of the intervention

Studies varied in their particular focus but, for the purposes of this review, understanding of science ideas was the central concern. In most cases, even if the authors had several variables they wished to study, these were investigated in such a manner as to obtain data relevant for this review.

In a number of studies, scientific understanding (science knowledge and/or scientific approach) was the sole and explicit variable of interest: Akpan and Andre (2000), Chang (2000), Dimitrov et al. (2002), Huppert et al. (2002) and Miglino et al. (2004).

In two of his studies, Chang (2001a, 2003) used the multimedia software resources that he had developed to answer different pedagogical questions; in the first investigation, he looked at a problem-solving approach, while in the later he was interested in the relative benefits of teacher-directed or student-directed, self-paced study. Huffman et al. (2003) were principally interested in two levels of ICT experience of the teachers but included non-ICT teachers and classes for a control, which provides relevant information for this review.

The focus of the intervention in the study by Diehl (2000) was scientific understanding (the scientific approach and argumentation) and also on beliefs and attitudes towards science.

Appropriateness of the measures

In most studies, the measures used were largely appropriate, although there was a common tendency to rely solely on multiple-choice questions for their ease of administration. Akpan and Andre (2000) used multiple-choice and short-answer tests.

The assessments of science understanding were either relevant established instruments or developed specifically for the evaluations.

Several studies used measures that discriminated levels of conceptual understanding. Chang (2000, 2001a, 2003) measured three conceptual levels: factual knowledge, comprehension, and application (ability to apply understanding gained in one context to another) as first proposed by Bloom (1956). Huppert et al. (2002) measured students' mastery of 'process skills' on a nine-point scale from simple ability to measure up to designing an experiment.

However, the understanding of scientific reasoning tested in the Diehl (2000) study combined two different types of information in the one assessment test. Questions on attitudes towards

Table 4.7
Weights of evidence C (WoE C) (N=9 studies)

	Study	WoE C
1	Akpan and Andre (2000)	MH
2	Chang (2000)	M
3	Chang (2001a)	MH
4	Chang (2003)	MH
5	Diehl (2000)	M
6	Dimitrov et al. (2002)	M
7	Huffman et al. (2003)	H
8	Huppert et al. (2002)	MH
9	Miglino et al. (2004)	M

Table 4.8
Overall weights of evidence (WoE D) assigned to each of the studies in the in-depth review (N=9 studies)

	Study	WoE A	WoE B	WoE C	WoE D
1	Akpan and Andre (2000)	ML	MH	MH	M
2	Chang (2000)	M	M	MH	M
3	Chang (2001a)	MH	MH	MH	MH
4	Chang (2003)	MH	MH	M	MH
5	Diehl (2000)	L	M	M	ML
6	Dimitrov et al. (2002)	MH	M	M	M
7	Huffman et al. (2003)	MH	MH	H	MH
8	Huppert et al. (2002)	M	M	MH	M
9	Miglino et al. (2004)	ML	ML	M	ML

Table 4.9
Summary of overall weights-of-evidence judgements (WoE D) on studies (N=9 studies)

Medium-high (MH)	Medium (M)	Medium-low (ML)
Huffman et al.(2003)	Akpan and Andre (2000)	Diehl (2000)
Chang (2001a)	Chang (2000)	Miglino et al. (2004)
Chang (2003)	Dimitrov et al. (2002)	
	Huppert et al. (2002)	

science – for example, the students' views of the importance of science outside school – were included in the same instrument as those on process, such as the role of evidence, hypothesis, and so on. The conflation of disparate data into one test makes the science achievement results of little value for this review.

Breadth

Seven studies reported on a broad range of science understanding: Chang (2000, 2001a, 2003), Dimitrov et al. (2002), Huffman et al. (2003), Huppert et al. (2002) and Miglino et al. (2004).

Akpan and Andre (2000) reported that they used the same test pre- and post the intervention. However the pre-test had 25 items, while the post-test had 43. There was no explanation for this difference. The Diehl (2000) study included a reasonable breadth of understanding about science process, such as evidence, hypotheses, explanations, contradictions and argument revisions. They did not set out to report on students' understanding of waste management, which was the science content of the intervention.

Situation

The studies by Diehl (2000), Huffman et al. (2003), and Huppert et al. (2002) involved students working in situations that were highly representative of learners in classrooms. That is, they were working singly or in small groups within the whole class.

The Akpan and Andre (2000) study reports that the students worked on the simulation individually and on the dissection in pairs, but the setting is not detailed. Students in the three Chang studies (2000, 2001a, 2003) spent much of their time working individually in a virtual research office which is not representative of how most classes operate.

Neither Dimitrov et al. (2002) nor Miglino et al. (2004) gave any information on how the students worked with the ICT.

Applying the above five criteria to each study produced the weights of evidence C (WoE C) displayed in Table 4.7.

An algorithm, previously used in Bennett et al. (2003, 2004), was used for combining weight of evidence judgements in categories A, B and C to arrive at an overall weight of evidence judgement in category D. This has been described in section 2.3.3 and an overview of the form may be found in Appendix 2.5. Application of the criteria and the algorithm give the judgements shown in Table 4.8.

Of the nine studies, three were rated medium-high (MH), four medium (M) and two medium-low (ML) as indicated in Table 4.19.

It was therefore considered appropriate to use the seven studies rated as medium high (MH) and medium (M) as the strongest basis for the in-depth review. The chief characteristics of these studies are that they are of a very good or good design and execution, and that their focus is particularly relevant to the review question.

The seven studies with medium-high (MH) and medium (M) chosen for the focus of this review were published by experienced researchers based at universities. None were graduate students or teacher researchers. Only one of these studies was a wholly independent evaluation and three included one or two independent researcher(s).

4.4 Synthesis of evidence

4.4.1 Summary findings of the seven studies

In the studies summarised below, effect sizes are presented in order to facilitate comparisons. Values of 0.8 or greater indicate large effects, 0.5 to just below 0.8 indicate medium effects and 0.20 to just below 0.5 small effects (Cohen, 1969). A number of the effect sizes were calculated by the authors, others by the reviewers. In one study, the reviewers were unable to calculate the effect size of the intervention as means and standard deviations data were not presented. Where results are referred to as 'significant' this refers to the statistical significance of the results as given by the study author.

Akpan and Andre (2000): BioLab frog dissection

Significant improvements were demonstrated in knowledge of frog anatomy and organ function with simulation-only (SO) and simulation-before-dissection (SBD) conditions. The improvements seen with dissection-only (DO) and dissection-before-simulation (DBS) were not significant. Effect sizes were 0.54 for SBD versus SO; 1.43 for SBD versus DBS, and 1.96 for SBD versus DO. (Reviewers' calculations of effect size are based on differences between pre- and post-achievement.)

WoE D: Medium

Chang (2000): Virtual office – animated weather satellite images and virtual field trip

Significantly higher student achievement scores were found for knowledge (effect size 0.20) and comprehension (effect size 0.12) levels compared with the score for students in traditionally taught (lecture from the teacher with some class discussion) classes. There were no differences in performance with respect to students' ability to apply acquired knowledge to a new situation (application level) (effect size 0.01). (Reviewers' calculations of effect size are based on differences between pre- and post-achievement.)

(WoE D: Medium)

Chang (2001a): Virtual office – animated weather satellite images and virtual field trip

Significantly higher student achievement scores were demonstrated for knowledge level compared with students receiving a lecture-internet-discussion teaching approach (effect size 0.11). There were no significant differences in performance with respect to students' comprehension (effect size 0.13) or application levels (effect size –0.01). (Reviewers' calculations of effect size are based on differences between pre- and post-achievement.)

WoE D: Medium-high

Chang (2003): Virtual office – animated weather satellite images and virtual field trip

Students experiencing the teacher-directed, multimedia software, including simulations, scored significantly higher than those in the student-directed, self-paced group using the same multimedia software at the knowledge (effect size 0.20) and application level (effect size 0.21), but not at the comprehension level (effect size 0.00). (Author's calculations are used for effect size.)

(WoE D: Medium-high)

Dimitrov (2002): Virtual village (included library articles, lectures, experiments, databases, data-handling and a virtual press conference

This study sought to compare different levels of ICT provision. All treatment students worked within the virtual environment with the same content. Participating teachers in each of the three experimental groups recruited another teacher at their school for the no treatment control. However, the two experimental groups (Search for Life and Mission to Pluto) had access to image analysis facilities but the alternative group did not. Significant effects were found for scientific knowledge understanding and for problem-solving (scientific approach) for all three treatments. For knowledge understanding and for problem-solving, the greatest change was with the Search for Life, then the alternative treatment, followed by the Mission to Pluto. The comparison of treatment effects across scales shows that the Search for Life group and the alternative group gained more in knowledge understanding than in problem-solving. The opposite was true for the Mission to Pluto group. (It was not possible to calculate the effect size from data presented on logits ability scale.)

WoE D: Medium

Huffman et al. (2003): Simulation of physics experiments on motion and force

Comparing three levels of teacher experience showed that students in the 'lead'/experienced teachers' classes had the highest science understanding scores, the students in the 'beginning' teachers' classes had the next highest, and the students in the comparison/traditional teaching classes had the lowest scores. The Cohen effect size differences in gain scores were 1.08 between the 'lead'/experienced teachers and the comparison/traditional teachers; 0.70 between the 'beginner' teachers and the comparison/traditional teachers; and 0.47 between 'lead/experienced teachers and 'beginning' teachers. (Authors' calculations of effect size is based on differences between pre- and post-achievement.)

WoE D: Medium-high

Huppert et al. (2003): Virtual experiment with growth of micro-organisms

Data was gathered on students' knowledge (of the population growth rate of micro-organisms) and nine science process skills in relation to the students' assessed cognitive stages as reasoners for both experimental and control groups, and for gender. The cognitive stages were (from lowest to highest) concrete, transitional and formal. This gave 10 x 3 x 2 x 2 variables that were analysed in a number of combinations. The study found that for scientific knowledge, both concrete and transitional, reasoners did better with the simulations than without (effect sizes of 2.66 and 2.83 respectively). There was no difference for formal (more advanced) reasoners between the experimental and control group (effect size 1.05). When achievement was analysed within groups, it was found that for both groups the transitional and formal groups did better than the concrete groups. For process skills, experimental groups did better than the controls for concrete and transitional groups in respect of the lower end skills of measurement, graph communication and interpreting data (effect sizes ranging from 1.73 to 8.12). Formal thinkers showed no differences for any of the nine skills between those using the simulations and those not (effect sizes ranging from 0.25 to 1.47). (Effect sizes were calculated by the authors.)

WoE D: Medium

4.4.2 Synthesis

As study numbers are low, studies are diverse in their uses of simulation activities and they vary in their research aims, only one or two examples are available in support of some of the following conclusions. Similarly, because the studies cover three different science subjects, were carried out in different countries (and hence educational systems), and used different types of simulation, no statistical meta-analysis is carried out. (The ranges of effect sizes are quoted in a and b, where several studies' findings are described together. Effect size details for individual studies discussed in c to g are also given.)

A synthesis of the findings of the studies in the in-depth review led to the following seven conclusions:

a. Use of ICT simulations helped significantly to improve students' understanding of science ideas compared with the use of non-ICT teaching activities.

Six studies allow for such a comparison between the learning effect of ICT simulations and non-ICT activities for the same intended learning outcomes (Akpan and Andre, 2000; Chang, 2000; Chang, 2001a; Dimitrov et al., 2002; Huffman et al., 2003; Huppert et al., 2003). All six studies support this conclusion. Effect sizes for statistically significant data ranged from 0.11 to 8.12. The seventh study (Chang, 2003) used simulation in the experimental and control conditions, but varied the teacher contribution.

b. Students understand science ideas significantly better when using ICT simulations versus their use of traditional (non-ICT) activities. This can apply to understanding of science knowledge (based on seven studies) and to understanding of the scientific approach (three studies).

The studies on which this conclusion is based are listed in Tables 4.10a and b.

c. The positive effect of students' use of ICT simulations on their understanding of science ideas is independent of the type of simulation, that is, simulations as virtual experiments (four studies) or simulations of a virtual environment (three studies).

The simulations fell into two main categories: (i) simulation of specific experiments and (ii) simulations of a wider scientific situation, commonly known as 'virtual environments', which could include experimental simulations.

Simulations of experiments
The three virtual experiments (Akpan and Andre, 2000; Huffman et al., 2003; Huppert et al., 2002) offered a number of advantages over traditional and real-time scientific teaching and learning. Dissection of real animals, for example, is less acceptable to some individuals than previously (Akpan and Andre, 2000) and the growth of micro-organisms can take days in the laboratory but a 'short time' on the computer (Huppert et al., 2002). Effect sizes ranged from 0.54 to 8.12.

Virtual environments
The Astronomy Village provided students across the USA with 'a virtual mentor guide' that helps students to complete multiple investigation cycles in the virtual village that mirror the phases of scientific enquiry about life, earth and physical science (Dimitrov et al., 2002). Chang's (2000, 2001a, 2003) studies provided a full range of resources, including a virtual field trip, in the virtual office within a classroom. Effect sizes ranged from 0.12 to 0.20.

d. Students' use of ICT simulations was more effective than using non-ICT teaching activities for supporting *basic* science ideas (from three studies) including the improvement of:

- Bloom's lower levels of understanding (two studies)
- understanding of basic aspects of the scientific approach (one study)
- science knowledge of less advanced reasoners (one study)

Chang (2000, 2001a) found that the virtual environment he provided gave an advantage to his students in the more basic knowledge (learning facts) and comprehension (understanding concepts) measures, but gave no benefit in the more difficult 'application' measure (applying acquired knowledge to a new situation).

Only one study (Huppert et al., 2002) set out to consider the students' cognitive level (along with a number of other variables) as part of their study. Within group analysis, it was found that for scientific knowledge both concrete (lower cognitive stage) and transitional reasoners did better with the simulations than without. There was no difference for formal (more advanced) reasoners between the experimental and control groups. Comparison of achievement between groups showed that both the transitional and formal groups did better than the concrete groups. For the scientific approach ('process skills'), experimental groups did better than the controls for concrete and transitional groups in respect of the lower-end skills of measurement, graph communication and interpreting data. Formal thinkers showed no differences for any of the nine skills between those using the simulations and those not. Thus the simulations gave some advantage to the less advanced reasoners with respect to scientific approach.

e. The improvements in higher understanding (for example, application), of more advanced aspects of the scientific approach (for example, the design of an experiment) and for more advanced (formal) reasoners can be achieved to the same extent with or without simulations.

The studies are described above in (d).

f. The gains from the students' use of ICT simulations were even further increased when teachers actively scaffolded or guided students through the ICT simulations (two studies).

Teacher as guide
Chang (2003) pointed explicitly to the importance of structured guidance with the use of the simulations, in this case through teacher support. Students who had access to the same material and ICT facilities, but worked individually and were self-directed and self-paced (SCCAI) achieved less than those who experienced the same content and multimedia resources as a whole class and were directed by a teacher (TDCAI). In both cases, students worked interactively. He suggests that it 'may be because the TDCAI provided students with systematic instructional content and organized teaching sequences, which may have facilitated students' understanding … and helped them grasp scientific facts and concepts' (p 435). Also 'the teacher-centred teaching method makes clear to the student what the objectives are and clarifies which learning materials and information are the most important. On the other hand, the SCCAI has probably left students with a large amount of data and information, which might have impeded students' learning of structured knowledge' (p 435).

Table 4.10a
Studies in which scientific understanding was significantly better with ICT simulations than without

Simulation activity	Scientific knowledge	Study
Dissection	Frog anatomy and organ function	Akpan and Andre (2000)
Virtual environment	Damage caused by typhoons	Chang (2000, 2001a, 2003)
Virtual environment	Astronomy – 'Mission to Pluto' and 'Search for Life'	Dimitrov et al. (2002)
Simulated experiments	Motion and forces	Huffman et al. (2003)
Simulated experiments	Population growth of micro-organisms	Huppert et al. (2003)

Effect sizes for statistically significant data ranged from 0.11 to 2.83.

Table 4.10b
Studies in which scientific approach was significantly better with ICT than without

Simulation activity	Scientific approach	Study
Virtual environment	Application of acquired knowledge to new situation	Chang (2003)
Virtual environment	Thinking and problem-solving; the transfer of thinking and PS skills	Dimitrov et al. (2002)
Simulated experiments	Science process skills	Huppert et al. (2003)

Effect sizes for statistically significant data ranged from 0.21 to 8.12.

Teaching style
Huffman et al. (2000) recognised that using ICT can involve a different teaching and learning style from that of the traditional, didactic approach. It requires that a constructivist approach be adopted that allows students to start from their own point of understanding as they work individually or in small groups. Thus the teacher must be familiar with that style of teaching as well as the ICT in use. Their study showed that not only did the ICT improve students' science understanding but that it could be even further enhanced when the teachers were experienced in the Constructing Physics Understanding ICT and constructivist approach.

A further indication of the need for structured guidance when using ICT simulations is the emphasis by a number of researchers on the guidance provided within the ICT simulation (Chang, 2000, 2001a, 2003; Dimitrov et al., 2002; Huppert et al., 2002). However, this aspect was not the specific focus of their evaluations.

g. The extra gains resulting from teacher guidance through the ICT simulations included further improvement of lower levels of understanding of science knowledge and approach, including the application of science knowledge to new situations.

Details are given in (f).

Thus simulations can bring benefits to students in respect of their understanding of science knowledge and scientific approach, but not in all situations and with all students and teachers. Care needs to be taken in establishing the particular benefits for particular learners and learning objectives in particular situations.

4.5 In-depth review: quality-assurance results

The quality-assurance processes for in-depth reviewing described in section 2.3.5 were followed. No areas of significant disagreement between the pairs of experts remained after moderating the data-extraction summaries. Generally, guidelines provided by collaborators at the EPPI-Centre were followed. The algorithm for determining the weighting of evidence of categories B and C worked well in securing coherence of these judgements across data-extraction teams. Additionally, the three members of the core team 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.

4.6 Nature of actual involvement of users in the review and its impact

The first meeting of the EPPI-Centre Review Group for Science, held on 7 January 2003, identified the effectiveness of ICT as one of the priority areas for a review of research findings. Members of the Training Development Agency for Schools (England and Wales) and the EPPI-Centre Review Group for Science were consulted when the review question and the in-depth review question were chosen.

The review findings will form the focus of a course session with ITT students when they have completed their teaching placements.

CHAPTER FIVE

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:

> *What is the effect of using ICT teaching activities in science lessons on students' understanding of science?*

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

> *What evidence is there from controlled trials of the effects of simulations on the understanding of science ideas demonstrated by students aged 11-16?*

5.1.2 Mapping of all included studies

Thirty-seven studies met the inclusion criteria developed for the overall research review. These studies were keyworded and formed the basis of the systematic map. The map revealed a number of characteristics of research on the use of ICT in science education, as summarised below:

- The majority of the studies reported work that had taken place in the USA (43%) and Taiwan (22%).
- A little over one-third of the studies concerned Biology topics and just under one-third were on Physics topics. Very little research has been done in relation to Chemistry education.
- Few authors gave explicit details of the ability range of their participant students. (It was therefore assumed that students were mixed ability or average for their age, unless otherwise stated.)
- In one-third of the studies, students worked individually with the ICT and in eight studies (22%), students worked in pairs. Two-fifths of authors (15 studies) did not give details of how the students interacted with the computers.
- Close to 90% of the studies focused on the students' understanding in respect of scientific knowledge/explanations and half on scientific approach; 12 studies investigated both. This interest was spread across Earth Science, Biology and Physics.
- Types of ICT activities used varied but half were referred to as simulations, either of experiments or of virtual environments. Virtual environments included a range of other ICT activities and non-ICT resources and could be defined as multimedia. Thus there is some overlap and flexibility in how the various forms of ICT activities are described or named.
- Fifty percent of the studies were carried out in one school with several classes. Only four studies (11%) involved large samples over several schools. Nine studies did not give full details of how many schools or classes were involved, although they all gave student numbers.
- Three-quarters of the studies used pre-post testing and half used questionnaires. Test results (that is, post- but no pre-test) were used in a quarter of the studies, as were interviews. Eight studies (22%) observed the student activities.
- Three-quarters of the studies were published in academic journals, seven (19%) as conference papers and one in a book chapter.
- Simulations/virtual environments, multimedia and moving images were used in the same or similar proportions to teach both scientific knowledge and scientific approach. (This is not too surprising given the overlap in these three

Table 5.1
Overall weights of evidence assigned to studies

Overall weight of evidence	Number of studies	Study
High	-	-
Medium-high	3	Huffman et al. (2003) Chang (2001a) Chang (2003)
Medium	4	Akpan and Andre (2000) Chang (2000) Dimitrov et al. (2002) Huppert et al. (2002)
Medium-low	2	Diehl (2000) Miglino et al. (2004)
Low	-	-

Table 5.2a
Studies investigating the impact of simulation activities on scientific knowledge

Simulation activity	Scientific knowledge	Study
Dissection	Frog anatomy and organ function	Akpan and Andre (2000)
Virtual environment	Damage caused by typhoons	Chang (2000, 2001a, 2003)
Virtual environment	Astronomy – 'Mission to Pluto' and 'Search for Life'	Dimitrov et al. (2002)
Simulated experiments	Motion and forces	Huffman et al. (2003)
Simulated experiments	Population growth of micro-organisms	Huppert et al. (2003)

Table 5.2b
Studies investigating the impact of simulation activity on understanding of scientific approach

Simulation activity	Scientific approach	Study
Virtual environment	Application of acquired knowledge to new situation	Chang (2000, 2001a, 2003)
Virtual environment	Thinking and problem-solving; the transfer of thinking and PS skills	Dimitrov et al. (2002)
Simulated experiments	Science process skills	Huppert et al. (2003)

Akpan and Andre (2000) did measure dissection performance (process skill) but did not report the finding in their study.

ICT categories.) Data-logging, databases, the internet and tutorial applications were used more often to teach scientific approach than scientific knowledge.

5.1.3 Nature of the studies selected for in-depth review

Care was taken to focus the in-depth review on those studies that most closely fulfilled the objectives of the review question:

- Which type of ICT activity was mostly evaluated in the research reports?
- Which were the most appropriate study designs for an evaluation (researcher manipulated / controlled)?
- Which studies involved a pre-/post-design?
- Which studies involved representative or average/typical students?

The aspect of ICT most frequently evaluated proved to be simulations (19 of the 37 studies).

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

The two studies with Medium-low weight of evidence were not used in the synthesis. This left seven studies which were relevant for this review and of a sufficient quality (medium-high and medium) from which to draw conclusions.

Tables 5.2a and 5.2b summarise the type of simulation activity and the type of science understanding being investigated.

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

The synthesis led to the seven conclusions listed below. However, as the sample size was small, the simulations variable, the learning objectives diverse and some of the following observations are based on only one study, a number of these findings should not be used for generalisations.

1. Students' use of ICT simulations helped to improve their understanding of science ideas significantly more effectively compared with their use of non-ICT teaching activities (based on six studies).
2. Students' significantly better understanding of science ideas when using ICT simulations versus their use of traditional (non-ICT) activities can lead to understanding of science knowledge (based on seven studies) and to understanding of scientific approach (three studies).
3. The simulations fell into two main categories: simulation of specific experiments and simulations of a wider scientific situation, commonly known as virtual environments, which could include experimental simulations.
4. The positive effect of students' use of ICT simulations on their understanding of science ideas is independent of the type of simulation, that is, simulations as virtual experiments (four studies) or simulations of a virtual environment (three studies).
5. Students' use of ICT simulations was more effective than using non-ICT teaching activities for supporting *basic* science ideas (from three studies), including the improvement of:
   - Bloom's lower levels of understanding (two studies)
   - understanding of basic aspects of the scientific approach (one study)
   - science knowledge of less advanced reasoners (one study).
6. The improvements in higher understanding (for example, application), of more advanced aspects of the scientific approach (for example, the design of an experiment) and for more advanced (formal) reasoners can be achieved to the same extent with or without simulations.
7. The gains from the students' use of ICT simulations were even further increased when teachers actively scaffolded or guided students through the ICT simulations (two studies). The extra gains resulting from teacher guidance through the ICT simulation included further improvement of lower levels of understanding of science (knowledge) and of the scientific approach, including the application of science knowledge to new situations (two studies).

Thus simulations can bring benefits to students in respect of scientific knowledge/explanations and approach, but not in all situations and with all students and teachers. Care needs to be taken in establishing the particular benefits for particular learners and learning objectives in particular situations.

5.2 Strengths and limitations of this systematic review

Strengths

The review has a number of strengths:

- The focus is one that is very relevant to the increased use of ICT in science teaching and learning. In particular, simulations are shown to be used in a wide range of situations. Evidence for this comes from the review map where 19 out of the 37 studies (51%) have simulations as their core mode of ICT.
- The evaluation studies considered student achievement in the spheres of scientific understanding and scientific approach.
- The approach to the review set high standards for the in-depth sample as only evaluation studies that had a control and pre-post test design were included. Additionally, the review only involved those studies that ensured their measures and their methods of analysis were valid and reliable.
- Quality-assurance results are high for all stages of the review.

Limitations

There are four main limitations:

- Although 19 evaluation studies involving simulations were found for this review, only seven were of a sufficient standard to include for the synthesis. These can thus only present successful examples of possibilities for teaching and learning in science education and highlight pedagogical points for consideration when using simulations; generalisations cannot be made.
- Some of the terms used in the field of ICT and education appear to be rather fluid. Thus model/modelling/a model can be used in the sense of 'to mimic or represent' or could mean to provide a predictive facility or process. Similarly,

simulation can be used to mean 'something has been modelled'. In this review, the predictive and more mathematical use of modelling was not included. (It did not anyway feature as a topic for evaluation studies.) Multimedia can also include simulations, in which case it is necessary to tease out the particular contribution of the simulation to learning effect.

- The in-depth studies covered the subjects of Earth Science, Biology and Physics. Only two studies out of the 37 in the map and none in the in-depth sample were in Chemistry. There appears to be a gap in this area of research which has impacted on this review.

- None of the in-depth studies was carried out in UK schools and thus the findings might not be directly applicable to the British educational situation. However, the fact that similar in-depth findings did emerge from three different countries (USA, Taiwan and Israel) does suggest that there is a measure of robustness in the findings that would make them of use in the United Kingdom.

5.3 Implications

5.3.1 Policy

Evaluation studies have found that ICT, and simulation in particular, can be helpful in teaching science understanding in respect of both scientific knowledge and scientific approach. However, it should be noted that there is a scarcity of high quality research in the area in which the in-depth study focused.

Teachers will also need training in the use of the simulations to obtain the greater benefit for student understanding. In particular, this review has shown that the use of ICT simulation needs to be carefully integrated into the teaching and learning process, and informed guidance provided. This guidance may be built into the software so that the students may work semi-independently or it may be provided by the teacher. However, teacher guidance is the more effective. This has implications for policies for initial teacher training and continuing professional development.

5.3.2 Practice

The review has indicated that there is a lack of clarity in the way that ICT and especially simulations, models and multimedia are being interpreted. One implication for practice is that teachers should be made aware of this.

The development of ICT simulations for a large variety of virtual experiments and virtual environments would provide a number of teaching and learning benefits. These include, *inter alia*, saving experimental time and resources, reducing the need to kill animals for dissection, allowing students to repeat experiments with ease, and providing experiences (through virtual environments) that would not otherwise be available to students.

The importance of the structured or guided use of ICT in particular simulations needs to be stressed to teachers. It is not sufficient just to provide the software unless it has in-built guidance or a virtual mentor. If it has not, the teacher needs to provide that support. Teachers may also need induction or training if the simulation is part of a complex teaching programme.

The inclusion of simulation activities within science Post Graduate Certificate in Education (PGCE) programmes would also encourage their use.

The newly established Regional Science Learning Centre could provide ideal opportunities for continuing professional development in the use of ICT in science education.

5.3.3 Research

The low numbers of high quality research studies into the value of using ICT in science education, especially in Chemistry, was surprising given the potential benefits. The use of ICT is likely to increase rather than decrease in schools in the near future. It is also likely that curriculum developers and commercial enterprises will increasingly develop software packages for science education.

It would therefore be of significant advantage if any science education ICT, of whatever origin, is carefully evaluated before it is adopted.

It is therefore suggested that more research is needed into the following:

- comparing the effect of different ICT teaching activities (for example, simulations with moving images)

- the nature of effective teacher support for ICT simulation activities.

CHAPTER SIX

References

6.1 Studies included in map and synthesis

*Studies in **bold** were included in the in-depth review.*

Linked papers in map are indicated with an asterisk.*

Akpan J (2001) Issues associated with inserting computer simulations into biology instruction: a review of the literature. *Electronic Journal of Science Education* 5(3): http://unr.edu/homepage/crowther/ejse/akpan.html.

Akpan JP, Andre T (2000) Using a computer simulation before dissection to help students learn anatomy. *Journal of Computers in Mathematics and Science Teaching* 19: 297-313.

Ates S, Stevens J (2003) Teaching line graphs to tenth grade students having different cognitive developmental levels by using two different instructional modules. *Research in Science and Technological Education* **21**: 55-66.

Atkinson S (2004) A comparison of pupil learning and achievement in computer aided learning and traditionally taught situations with special reference to cognitive style and gender issues. *Educational Psychology* **24**: 659-679.

Bakas C, Mikropoulos TA (2003) Design of virtual environments for the comprehension of planetary phenomena based on students' ideas. *International Journal of Science Education* **25**: 949-967.

Chang C-Y (2000) Enhancing tenth graders' earth-science learning through computer-assisted instruction. *Journal of Geoscience Education* 48: 636-640.

Chang C-Y (2001a) A problem-solving based computer-assisted tutorial for the earth sciences. *Journal of Computer Assisted Learning* 17: 263-274.

*Chang C-Y (2001b) Comparing the impacts of a problem-based computer-assisted instruction and the direct-interactive teaching method on student science achievement. *Journal of Science Education and Technology* **10:** 147-153.

*Chang C-Y (2002a) Does computer-assisted instruction + problem solving = improved science outcomes? A pioneer study. *Journal of Educational Research* **95:** 143-150.

*Chang C-Y (2002b) The impact of different forms of multimedia CAI on students' science achievement. *Innovations in Education and Teaching International* **39:** 280-288.

Chang C-Y (2003) Teaching earth sciences: should we implement teacher-directed or student-controlled CAI in the secondary classroom? *International Journal of Science Education* 25: 427-438.

Chang C-Y (2004) Could a laptop computer plus the liquid crystal display projector amount to improved multimedia geoscience instruction? *Journal of Computer Assisted Learning* **20**: 4-10.

Chang KE, Sung YT, Chen SF (2001) Learning through computer-based concept mapping with scaffolding aid. *Journal of Computer Assisted Learning* **17**: 21-33.

ChanLin L-J (2001) Effects of presentation formats and prior knowledge on learning in a computer-based lesson. *Journal of Computer Assisted Learning* **17**: 409-419.

Choi E, Park J (2003) Conditions for the effective use of simulation and its application to middle-school physics inquiry activities. *Journal of the Korean Physical Society* **42**: 318-324.

Cifuentes L, Yi-Chuan JH (2000). *Concept learning through image processing*. In: Annual Proceedings of Selected Research and Development: Papers presented at the National Convention of the Association for Educational Communications and Technology, Denver, CO, Volumes 1-2.

Clark DB, Slotta JD (2000) Evaluating media-enhancement and source authority on the internet: the knowledge integration environment. *International Journal of Science Education* **22**: 859-871.

Diehl CL (2000) '*Reasoner's Workbench' program supports students' individual and collaborative argumentation*. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, New Orleans, LA.

Dimitrov DM, McGee S, Howard BC (2002) Changes in students' science ability produced by multimedia learning environments: application of the linear logistic model for change. *School Science and Mathematics* 102: 15-24.

Hennessy S (2000) Graphing investigations using portable (palmtop) technology. *Journal of Computer Assisted Learning* **16**: 243-258.

Hoadley CM, Linn MC (2000) Teaching science through online, peer discussions: SpeakEasy in the knowledge integration environment. *International Journal of Science Education* **22**: 839-857.

Hong NS, McGee S, Howard BC (2000). *The effect of multimedia learning environments on well-structured and ill-structured problem-solving skills*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Hsu Y-S, Liao K-C, Chen C-C (2001) *A web-based lesson with situated learning in senior high school level*. In: WebNet 2001 World Conference on the WWW and Internet Proceedings, Orlando, FL.

*Hsu Y-S, Hwang F-K (2002). *The use of multiple representations in a web-based and situated learning environment*. In: ED-MEDIA 2002 World Conference on Educational Multimedia, Hypermedia and Telecommunications Proceedings, Denver, Colorado.

Huffman D, Goldberg F, Michlin M (2003) Using computers to create constructivist learning environments: impact on pedagogy and achievement. *Journal of Computers in Mathematics and Science Teaching* 22: 151-168.

Huppert J, Lomask SM, Lazarowitz R (2002) Computer simulations in the high school: students' cognitive stages, science process skills and academic achievement in microbiology. *International Journal of Science Education* 24: 803-821.

Kariuki P, Paulson R (2001) *The effects of computer animated dissection versus preserved animal dissection on the student achievement in a high school biology class*. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Little Rock, AR.

*Koroghlanian C, Klein JD (2000). *The use of audio and animation in computer based instruction*. In: Annual Proceedings of Selected Research and Development Papers Presented at the National Convention of the Association for Educational Communications and Technology, Denver, CO, Volumes 1-2.

Koroghlanian C, Klein JD (2004) The effect of audio and animation in multimedia instruction. *Journal of Educational Multimedia and Hypermedia* **13**: 23-46.

Lin C-Y, Cheng Y-J, Chang Y-T, Hu R (2002) The use of internet-based learning in biology. *Innovations in Education and Teaching International* **39**: 237-242.

Marcum-Dietrich NI, Ford DJ (2002) The place for the computer is in the laboratory: an investigation of the effect of computer probeware on student learning. *Journal of Computers in Mathematics and Science Teaching* **21**: 361-379.

*McGee S, Howard BC, Dimitrov DM, Hong NS, Shia R (2002) *Addressing the complexities of evaluating interdisciplinary multimedia learning environments*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Means B, Coleman E (2000) Technology supports for student participation in science investigations. In: Kozma, RB, Jacobson, MJ (eds), *Innovations in Science and Mathematics Education: Advanced Designs for Technologies of Learning*. Mahwah, NJ: Lawrence Erlbaum Associates, pp 287-319.

Miglino O, Rubinacci F, Pagliarini L, Lund HH (2004) Using artificial life to teach evolutionary biology. *Cognitive Processing* 5: 123-129.

Mistler-Jackson M, Songer NB (2000) Student motivation and internet technology: are students empowered to learn science? *Journal of Research in Science Teaching* **37**: 459-479.

Ravenscroft A, Matheson MP (2002) Developing and evaluative dialogue games for collaborative e-learning. *Journal of Computer Assisted Learning* **18**: 93-101.

Ronen M, Eliahu M (2000) Simulation - a bridge between theory and reality: the case of electric circuits. *Journal of Computer Assisted Learning* **16**: 14-26.

Shim KC, Park JS, Kim HS, Kim JH, Park YC, Ryu HI (2003) Application of virtual reality technology in biology education. *Journal of Biological Education* **37**: 71-74.

Shimoda TA, White BY, Frederiksen JR (2002) Student goal orientation in learning inquiry skills with modifiable software advisors. *Science Education* **86**: 244-263.

Snir J, Smith CL, Raz G (2003) Linking phenomenon with competing underlying models: a software tool for introducing students to the particulate model of matter. *Science Education* **87**: 794-830.

Tao P-K (2000) *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.

Tsui C-Y, Treagust DF (2003) Genetics reasoning with multiple external representations. *Research in Science Education* **33**: 111-135.

Tsui C-Y, Treagust DF (2003) *Learning genetics with multiple representations: a three dimensional analysis of conceptual change*. Paper presented at the National Association for Research in Science Teaching, Philadelphia, PA.

Yildirim Z, Ozden MY, Aksu M (2001) Comparison of hypermedia learning and traditional instruction on knowledge acquisition and retention. *Journal of Educational Research* **94**: 207-214.

6.2 Other references used in the text of the report

Barton R (1997a) How do computers affect graphical interpretation? *School Science Review* **79**: 55-60.

Barton R (1997b) Does data-logging change the nature of children's thinking in experimental work in science? In:

Somekh B, Davies N (eds) *Using Information Technology Effectively in Teaching and Learning*. London: Routledge.

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

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.

Bennett J, Lubben F, Hogarth S, Campbell B, 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.

Bloom BS (1956) *Taxonomy of Educational Objectives. Handbook I: Cognitive Domain*. New York: David McKay.

British Educational Communications and Technology agency [Becta] (2001a) *Primary Schools of the Future: Achieving Today*. Coventry: Becta.

British Educational Communications and Technology agency [Becta] (2001b) *The Secondary School of the Future*. Coventry: Becta.

Campbell D, Stanley J (1996) *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Cohen J (1969) *Statistical Power Analysis for the Behavioural Sciences*. New York: Academic Press.

Collins J, Hammond M, Wellington, J (1997) *Teaching and Learning with Multimedia*. London: Routledge.

Department for Education and Employment/ Qualifications and Curriculum Authority [DfES/QCA] (1999) *Science: The National Curriculum for England, Key Stages 1-4*. London: DfES/QCA.

Department for Education and Skills [DfES] (2001) *Statistical Bulletin 36/2001*. London: The Stationery Office.

EPPI-Centre (2002a) *Core Keywording Strategy: Data Collection for a Register of Educational Research, Version 0.9.7*. London: EPPI-Centre, Social Science Research Unit.

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

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

Goldstein G (1997) *Information Technology In English Schools: A Commentary on Inspection Findings 1995-1996*. London: The Stationery Office.

Hennesey S, Twigger D, Driver R, O'Shea T, O'Malley C, Byard M, Draper S, Hartley R, Mohamed R, Scanlon E (1995) A classroom intervention using computer augmented curriculum for mechanics. *International Journal of Science Education* **17**: 189-206.

McKinsey and Co (1997) *The Future of Information Technology in UK Schools*. London: McKinsey and Co.

Nakhleh M, Krajcik J (1993) A protocol analysis of the influence of technology on students' actions, verbal commentary and thought processes during the performance of acid-base titrations. *Journal of Research in Science Teaching* **30**: 1149-1168.

National Council for Educational Technology [NCET] (1994) *Information Technology Works! Stimulate to Educate*. Coventry: NCET.

Newton L (2000) Data-logging in practical science: research and reality. *International Journal of Science Education* **22**: 1247-1259.

Office for Standards in Education [OfSTED] (2001) *ICT in Schools: The Impact of Government Initiatives*. London: OfSTED.

Office for Standards in Education [OfSTED] (2002) *The National Literacy Strategy: The First Four Years 1998-2002 Initiatives*. London: OfSTED.

O'Shea T, Scanlon E, Byard M, Draper S, Driver R, Hennessey S, Harteley R, O'Malley C, Mallen C, Mohammed G, Twigger D (1993) Twenty-nine children, five computers and a teacher. In: Edwards D, Scanlon E, West R (eds) *Teaching, Learning and Assessment in Science Education*. Buckingham: Open University Press.

Poole P (2000) Information and communications technology in science education: a long gestation. In: Sears J, Sorensen P (eds) *Issues In Science Teaching*. London: RoutledgeFalmer.

Raghaven K, Glaser R (1995) Model-based analysis and reasoning in science: the MARS curriculum. *Science Education* **79**: 37-61.

Rogers L, Newton N (2001) Integrated learning systems – an 'open' approach. *International Journal of Science Education* **23**: 405-422.

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

Vygotsky L S (1962) *Thought and Language*. Cambridge: MIT.

Vygotsky L S (1978) *Mind in Society*. Oxford: Blackwell.

Wellington J (1999) Integrating multimedia into science teaching: barriers and benefits. *School Science Review* **81**: 49-54.

Appendix 1.1: Review group membership

This work is a report of a systematic review conducted by the Review Group for Science

The authors of this report are

Sylvia Hogarth (University of York)
Judith Bennett (University of York)
Fred Lubben (University of York)
Bob Campbell (University of York)
Alison Robinson (University of York)

They conducted the review with the benefit of advice and active participation from the members of the review group.

For further information about this review, please contact:

Dr Sylvia Hogarth
Department of Educational Studies
University of York
York YO10 5DD
Tel: 01904 433475
Fax: 01904 433459
Email: sdh6@york.ac.uk

For further information about the work of the EPPI-Centre, please contact

EPPI-Centre
Social Science Research Unit
Institute of Education, University of London
18 Woburn Square
London WC1H 0NR
Tel: +44 (0)20 7612 6131
Fax: +44 (0)20 7612 6800
Email: EPPIAdmin@ioe.ac.uk

Review Group

Dr Judith Bennett (University of York)
Martin Braund (University of York)
Dr Bob Campbell (University of York)
Nick Daws (Ofsted and Staffordshire LEA)
Steve Dickens (Dixons City Technology College, Bradford)
Alison Fletcher (Huntington School, York)
Nichola Harper (Aldridge School, Walsall)
Dr Sylvia Hogarth (University of York)
Professor John Holman (University of York)
Declan Kennedy (University College, Cork)
Ralph Levinson (University of London)
Fred Lubben (University of York)
Alyson Middlemass (Elizabethan High School, Retford)
Professor Robin Millar (University of York)
Christine Otter (University of York and Salters Advanced Chemistry)
Professor Mary Ratcliffe (University of Southampton)
Alison Robinson (University of York)
Daniel Sandford-Smith (Institute of Physics)
Dr Carole Torgerson (University of York)

Conflicts of interest

There are no conflicts of interest associated with this review

Acknowledgements

The Science Review Group is one of a number of groups undertaking systematic reviews to support the evidence base of initial teacher training (ITT) staff in England.

The Review Group is grateful for the funding from the Training Development Agency for Schools (TDA) which enabled this review to be carried out. The Review Group also acknowledges the financial support via core institutional funding from the Higher Education Funding Council for England (HEFCE). The Review Group is grateful to the EPPI-Centre at the Institute of Education, University of London, for its support for the review, and for permission to use its systematic review tools.

Appendix 2.1: Inclusion and exclusion criteria

Review question:

What is the effect of using ICT teaching activities in science teaching on students' understanding of science ideas?

Inclusion criteria

- Must be a study of ICT in science teaching
- Must be a study of the effects of using ICT teaching activities on understanding of science ideas
- Must focus on students aged 11-16
- Must be in a mainstream school setting
- Must be an evaluation study

Exclusion criteria

Exclusion on scope

One Not Science learning/teaching

(Science: one or several of the school subjects, i.e. Integrated/General Science, Science, Biology, Chemistry, Physics or Earth Science. Not Mathematics, Technology, Social Science or Computing)

Two Not ICT teaching activities:

(ICT teaching activities: tutorial applications, simulations, modelling, data-logging, graphing, use of multimedia, use of the internet. Not word-processing of essays, assessment record-keeping using ICT)

Three Not effects on understanding of science ideas

Four Not children or young people of ages between 11 and 16.

Exclusion on study type

Five (a) A (description)

(b) B (exploration of relationships)

(c) D (methodology)

(d) E (review)

Exclusion on type of publication/source

Six (a) Editorial, commentary, book review, review

(b) Policy document

(c) Resource, text book

(d) Bibliography

(e) Dissertation abstract

(f) Theoretical paper

Exclusion on setting in which study was carried out

Seven Not mainstream school setting

Eight Not published in the period 2000–2004

In-depth exclusion criteria

One Did not focus on the use of simulations

Two Were not of a researcher-manipulated/controlled trial design

Three Did not report pre-post test results

Four Did not involve representative/average students

Appendix 2.2: Electronic search strategy

ERIC

1. KW=(teach* or learn*) and KW=(science or chemistry or biology or physics)
2. DE=(science education) or (science instruction)
3. 1 or 2
4. DE=(computer uses in education) or (computer assisted instruction) or (computer software) or (information technology) or (integrated learning systems) or (web based instruction) or (multimedia instruction)
5. DE=(secondary education) or (elementary secondary education) or (high school equivalency programs) or (elementary education) or (middle school*) or (secondary school*) or (high school*)
6. 3 and 4 and 5
7. Limit 6 to PT=(142 reports evaluative) or PT=(143 reports research) or PT=(080 journal articles) or PT=(010 books)
8. Limit 7 to (Language = English) and (Year = 2000-2004)

PsycINFO

1. (teach* or learn*) and (science or chemistry or biology or physics)
2. (science education) or (science instruction)
3. 1 or 2
4. computer* or software or ICT or CAI or CAL or internet or multimedia or (web based) or (information and communication technolog*) or (information technolog*) or (digital technolog*)
5. (secondary education) or (secondary school*) or (middle school*) or (high school*)
6. 3 and 4 and 5
7. Limit 6 to (DT = authored-book) or (DT = chapter) or (DT = edited-book) or (DT = journal) or (DT = peer-reviewed-journal) or (DT = report)
8. Limit 7 to (Language = English) and (Year = 2000-2004)

SSCI

1. TS=(teach* or learn*) and (science or chemistry or biology or physics)
2. TS=(science education) or (science instruction)
3. 1 or 2
4. TS=(((secondary education) or (secondary school*) or (middle school*) or (high school*)) and (computer* or software or ICT or CAI or CAL or internet or multimedia or (web based) or (information and communication technolog*) or (information technolog*) or (digital technolog*)))
5. 3 and 4
6. Limit 5 to (Language=English) and (Year = 2000-2004)

BEI

1. KW=(teach? or learn?) and (science or chemistry or biology or physics)
2. KW=(science education) or (science instruction)
3. 1 OR 2
4. DE=('computer assisted learning') or ('computer uses in education')
5. KW=(information technolog?) or (information and communication technolog?) or (networked technolog?) or (digital technolog?) or (multimedia instruction) or (computer? or software or ICT or CAI or CAL)
6. 4 or 5
7. KW=(secondary education) or (secondary school?) or (middle school?) or (high school?)
8. 3 and 6 and 7
9. Limit 8 to (Language = English) and (Year = 2000-2004)

APPENDIX 2.3 EPPI-Centre keyword sheet, including review-specific keywords

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

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

A2. Status
Published
In press
Unpublished

A3. Linked reports
Is this report linked to one or more other reports in such a way that they also report the same study?
Not linked
Linked (please provide bibliographical details and/or unique identifier)
. . . .
. . . .
. . . .
. . . .

A4. Language (please specify)
. . . .

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

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

A7. Curriculum
Art
Business studies
Citizenship
Cross-curricular
Design and technology
Environment
General
Geography
Hidden
History
ICT
Literacy - first language
Literacy further languages
Literature
Mathematics
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

Appendix 2.4: Review-specific keywords

1. What discipline?

a. Integrated Science
b. Biology
c. Chemistry
d. Physics
e. Earth Science
f. Other ....

2. What types of learners are involved?

a. Mixed ability
b. Lower ability/slow learners
c. Upper ability/gifted
d. Disaffected
e. Unspecified
f. Other ....

3. How do students work with ICT?

a. Singly in school
b. In pairs in school
c. In groups up to 5 in school
e. In groups of 5 or more in school
f. Whole class together in school
g. At home
h. Other...............................

4. On which aspect(s) of understanding of science ideas does the study focus? (Please tick main aspect(s) only.)

a. Scientific knowledge/explanations (facts, concepts, laws, theories)
b. Scientific approach (evidence, scientific methods, problem solving)
c. Ideas about science (limitations, scientific community, risk, etc.)
d. Applications of science
e. Other...

5. On which aspect(s) of ICT does the study focus? (Please tick main aspect(s) only.)

a. CAI/CAL (no details on type)
b. Modelling
c. Simulations/virtual environments
d. Data-logging
e. Use of databases
f. Multimedia
g. Internet/WWW
h. Online discussion
i. Tutorial applications
j. Hypertext
k. Moving image (animations, video clips)
l. Email
m. Mobile phone
n. Games/adventures
o. Other...

6. Scope of study

a. Science understanding is the only focus
b. Science understanding one of several aspects studied
c. Other...

7. Size of study

a. Across several/many schools
b. One school, several classes
c. One class
d. Small group of students
e. Other.......................................

8. Which outcomes are reported?

a. Pre- and post-test results
b. Test results
c. Examination results
d. Written reports /open questionnaires
e. Computer logs /computer products
f. Observed behaviour (including video)
g. Recorded group discussions (audio)
h. Interviews
i. Work diaries/logs
j. Presentations
k. (Dis)agreement scores
j. Other..

9. Type (quality) of paper

b. Conference paper
a. Academic journal
c. Teacher journal
d. Book chapter
e. Other...

10. Variables

a. ICT is only variable.
b. ICT is main variable.

APPENDIX 2.5 Indicators for weight of evidence

Review question: What evidence is there from controlled trials of the effects of computer simulations on the understanding of science ideas demonstrated by students aged 11-16

WEIGHT OF EVIDENCE B:

Appropriateness of research design and analysis for addressing the question *of this specific systematic review*

high (3)	medium (2)	low (1)
For the RQ relevant to the review		
Sample size		
Large sample with appr. sampling method	Large sample, no sampling method	Small sample (up to 3 classes)
Selection of control		
Random for students in intervention	Random for classes in intervention	No aspect of random control
Control of extraneous variables		
Deliberate attempt to control two or more factors other than simulation (e.g. gender, school, teacher, achievement)	Deliberate attempt to control less than three factors other than simulation (e.g. gender, teacher, school, achievement)	No attempt to control extraneous factors
Data collection		
Solid checks on rel/val for data collection	Some checks on rel/val for data collection	Little/no checks on rel/val for data collection
Data analysis		
Solid checks on rel/val for data analysis	Some checks on rel/val for data collection	Little/no checks on rel/val for data collection

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*

high (3)	medium (2)	low (1)
For the RQs relevant to the review		
Experience		
Students work together interactively on simulation	Students work singly on simulation	Students watching demonstration of simulation
Focus of intervention		
Understanding is sole and explicit independent variable	Understanding is major discrete element of intervention	Understanding is wrapped up in intervention
Measures		
Highly appropriate for testing directly understanding of science ideas	Mildly appropriate for testing directly understanding of science ideas	Appropriate for testing indirectly understanding of science ideas
Breadth		
Reports broad range of science understanding	Reports narrow range of science understanding	Only reports science understanding indirectly
Situation		
Highly representative of learners in classrooms	Less representative of learners in classrooms	Not in classrooms

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?*

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)

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

Akpan JP and Andre T (2000) Using a computer simulation before dissection to help students learn anatomy. *Journal of Computers in Mathematics and Science Teaching* 19: 297-313

Country of study
USA

Details of researchers
Based at two US universities

Name of programme (if applicable)
The dissection simulation was BioLab Frog software supplied by Pierian Spring. This incorporates QuickTime movies and microscopic pictures to illustrate functions that are normally hidden from view.

Age of learners
13-15 years

Type of study
Evaluation: researcher-manipulated

Aims of study
To examine whether computer simulations, used in conjunction with actual hands-on dissection, might lead to a better educational experience for students.

Summary of study design, including details of sample
Students were initially allocated to classes by the teacher in order to roughly equalise ability across classes.

Student 'class periods' were allocated randomly to the four study conditions giving a quasi-random design.

Students undertook one of four conditions: simulation before actual frog dissection; simulation after actual frog dissection; dissection only; simulation only

N = 81 (34 males and 47 females)

Methods used to collect data
A multiple-choice and short-answer test was used pre- and post-intervention.

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

- 25 questions were used for the pre-test and 43 questions for the post-test, and their reliability tested using Cronbach.
- Iowa test of basic skills (ITBS) was used as the covariate for the analysis.
- Checks on reliability: a preliminary Cronbach *a* analysis for internal consistency and post-test questions.
- Checks on validity: The test was designed by a classroom teacher with two science experts. However: 'the results are limited to the types of educational outcomes reflected in this test' and this may be inadequate in the light of calls for 'more authentic assessments'.

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

- Pre-test data: 2 gender x 4 Treatment ANCOVA with ITBS Science score as covariate
- Post-test data: 2 gender x 4 Treatment x 2 test time (pre-test vs. post-test) between/within ANCOVA with ITBS Science score as covariate
- Checks on reliability: None articulated but standard and appropriate test used
- Checks on validity: Use of pre-test established that there were no differences between the four conditions or two genders in respect of the covariate of student ability using ITBS as the covariate.
- Follow-up analyses of post-test data (gender x time between/within ANCOVAs) to confirm significance of result.
- As the pre- and post-test interval was not great, cognitive, development / maturational impact should be minimal.

Summary of results

- Pre-test: No significant main effects or interactions were found, indicating pre-experimental equivalence of the four conditions and the two genders.

- Post-test: All conditions showed improved post-test achievement and overall there was 'a significant main effect of condition, and of test time (i.e. pre-/ post-test difference but these main effects were modified by a significant test time by condition interaction).

- Two of the post-test improvements were significant: simulation-only and the simulation-before-dissection condition.

- The improvements in the remaining two conditions were not significant: dissection-only and dissection-before-simulation condition.

- No statistical differences with gender.

Conclusions

- The authors state 'The results of this study supported the theory that prior use of a simulation before dissection can improve learning. The treatment group that completed the simulation activities before the actual hands-on dissection performed significantly better on the achievement post-test and dissection performance test than the other three groups.' However, (1) inspection of the data and what they report in the results section shows that the simulation only (SO) group also made significant improvements. (2) The authors did not carry out a 'dissection performance test'; they did a post-intervention test.

- There was no difference in post-test achievement or dissection performance between the sexes in any condition.

- 'The most intriguing result of the present study was that a simulation used before dissection led to better achievement performance than a simulation after dissection… The results of this study suggest that presentation of a computer simulation before the actual dissection may provide an experiential base that enhances learning.'

- The results did not support the work of others who found gender differences; however, the other researcher who did find these differences was looking at simulations in the context of Physics activities.

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

Although the experimental design is satisfactory and data relevant to the study question is given, there are a number of concerns about the reporting of the study. Also all the conclusions reached in the discussion are not supported by this data: (1) Simulation before dissection is stated to be the best outcome although the data suggests that it is not significantly better than simulation only. (2) Data underlying the comparison of the four conditions is not given, only that for pre-test/post-test comparison within each. (3) No data is given for the conclusions regarding gender; dissection performance and attitudes are discussed in the conclusions, but no data is given in the results section.

Weight of evidence B (appropriateness of research design and analysis)
Medium-high

Small-medium sample (81 students/4 classes) although care was taken to ensure that students included had completed both the pre- and the post-tests; random for classes in the intervention; did rough control for ability and dissection experience; some checks for reliability and validity of data collection; used relevant statistical tests but did not present full data analysis findings.

Weight of evidence C (relevance of focus of study to review)
Medium-high

Students worked directly on the simulation but no details of whether singly or in pairs; understanding is major element of the intervention, gender also investigated; factual and short answer questions are appropriate for testing understanding; breadth of understanding was variable as 20 questions were used in the pre-test and 43 in the post-test. These covered anatomy and function; little information given about the class situation.

Weight of evidence D (overall weight of evidence)
Medium

Chang CY (2000) Enhancing tenth graders' earth-science learning through computer-assisted instruction. *Journal of Geoscience Education* 48: 636-640

Country of study
Taiwan

Details of researchers
Assumed to be university staff member and teacher

Name of programme (if applicable)
Not applicable

Age of learners
Mean age 16

Type of study
Researcher-manipulated evaluation

The author adopts the method of Campbell and Stanley (1966) which involves a pre-post, control group design. However, the author does not explicitly say that the allocation of the individuals to groups was random. Personal communication with the author established that he randomly assigns classes not individuals.

Aims of study
To evaluate the effectiveness of CAI methods as compared with traditional instruction methods for secondary schools in Earth Science.

Summary of study design, including details of sample
Four classes (total of 151 students) were allocated two each to the experimental group (N=79) with the use of CAI methods, and to the control group (N=72) with traditional teaching methods. The dependent variable is student concept achievement. No information of gender breakdown for both groups. Controlled for teacher, school administration, teaching time, curriculum content (p 638).

Methods used to collect data
Curriculum-based assessment: Pre-post tests of Earth Science achievement. 30 question, multiple choice test.

Data-collection instruments, including details

of checks on reliability and validity
- Used a self-constructed, written pre- and post-intervention test, the Earth Science Achievement Test (ESAT)
- Test comprises 30 multiple-choice items, 5 items cover knowledge, 18 items comprehension, 7 application achievement.
- High reliability of test. Allocation of items to knowledge, comprehension and application subgroups had high agreement for six experts
- Validity of the tests was checked using a panel of specialists, including three high school teachers and three professors.

Methods used to analyse data, including details of checks on reliability and validity
- Statistical (ANCOVA) analysis of covariance of pre- and post-test scores level of confidence set at $p<0.05$
- Use of accepted statistical methods.

Summary of results
- There were significant differences in students' overall learning of Earth Science concepts between subjects in the experimental group and the comparison group.
- Most notably, the CAI significantly improved student achievement compared with the traditional teaching method at the knowledge and comprehension levels.
- However, there was no statistical difference for application.

Conclusions
- The author states, 'The findings of this study show that CAI is superior in promoting students' learning of earth science concepts, especially knowledge and comprehension levels of Bloom's cognitive taxonomy.' This suggests that CAI could help students acquire knowledge and grasp geoscience concepts.
- The author speculates that the higher effectiveness of the CAI method may relate to students working on their own, and thus consider this way of learning more fun; the need for students to understand a central problem, gather information and formulate their own conclusions. The results are in line with previous empirical studies concluding that CAI is effective in teaching factual content.

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

The instrument has some drawbacks. It needs a bit more detail on gender and how allocation was made to experimental and control groups. The number of items for 'knowledge' achievement and for 'application' achievement are too few to be confident about the scores in these areas. The multiple-choice nature of the test is not optimal to measure comprehension and application. The main drawback is that nature of the interventions (what took place) remains unclear.

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

Large sample (151 students), but no sampling frame and no details of gender; not explicit that students were randomly assigned to control groups; controls established for school, teacher, age, group size and teaching time; solid checks for reliability and validity of data collection; appropriate data analysis test used, pre-post ANCOVA but no indication of effect size.

Weight of evidence C (relevance of focus of study to review)
Medium-high

Students work singly in CAI; achievement is explicit independent variable; multiple choice and few items for some sub-levels of understanding reduces the appropriateness of the instrument; good breadth of achievement of understanding; use of 'individual research office' not very representative of class work.

Weight of evidence D (overall weight of evidence)
Medium

1. Chang CY (2001a) A problem-solving based computer-assisted tutorial for the earth sciences. *Journal of Computer Assisted Learning* 17: 263-274

2. Chang CY (2001b) Comparing the impacts of a problem-based computer-assisted instruction and the direct-interactive teaching method on student science achievement. *Journal of Science Education and Technology* 10: 147-153

3. Chang CY (2002a) Does computer-assisted instruction + problem solving = improved science outcomes? A pioneer study. *Journal of Educational Research* 95: 143-150

Country of study
Taiwan

Details of researchers
Assumed to be university staff member and teacher

Name of programme (if applicable)
Not applicable

Age of learners
Mean age 16

Type of study
Researcher-manipulated evaluation

Aims of study
'To investigate the impacts of using a Problem-Solving based Computer-Assisted Tutorial (PSCAT) vs. a Lecture-Internet-Discussion (LID) teaching method on secondary-level students' earth science achievement'

Summary of study design, including details of sample
Pre-post tests; control group

Experimental group: problem-solving computer-assisted tutorial (PSCAT) (N=72); and control group: lecture-internet-discussion (LIC) (N=65); over one week

Four classes (rather than individuals) allocated randomly to the two treatments (personal communication from author) does not qualify as a full RCT. Treatments were controlled for the same teacher, school gender, ability (implicit) teaching content, teaching materials and assignments and teaching time.

N = 137 in 2001a paper and 159 in 2001b. However, these were the same students. The 137 sample is a subset of the 159 who had also completed the questionnaire on attitude.

Methods used to collect data
Curriculum-based assessment: multiple-choice test

Data-collection instruments, including details of checks on reliability and validity
No data to define the sample.

Use of Earth Science Achievement Test 30 item multiple-choice to measure: Knowledge (5 items); Comprehension (18 items); Application (7 items) – from Bloom's (1956) taxonomy. Checks on reliability: Allocation of items in Knowledge, Comprehension and Application categories by six experts, with high agreement (83%–90%).

High reliability coefficient (Kuder-Richardson Formula 20) of 0.76 to 0.78 for pre- and post-tests.

Checks on validity: Used a panel of experts to check for validity; three university professors from the Department of Earth Science and three high-school teachers.

Methods used to analyse data, including details of checks on reliability and validity
Statistical analyses: MANCOVA on achievement with pre-test measures as covariates. Wilks' lambda used for testing significance of differences between treatment groups for adjusted post-treatment results (significance level 0.05). ANCOVA on post-test achievement with pre-test as covariate on sub-level achievement (knowledge, comprehension, application) to determine significance of differences between treatment groups.

Checks on reliability: Used standard test with probability level of 0.05 applied for overall/total results. Also applied the Bonferroni method to control for Type 1 error rates for each of the sub-levels of analysis (knowledge, comprehension, application), so only accepted 0.017 as significance level.

Checks on validity: Used the Kolmogorov-Smirnof test to check for normal distribution of test results for both groups. Used Bartlett-Box F test for homogeneity of variances among groups. Assumption of homogeneity of regression of the covariate and the dependent variable was also examined and found to be tenable.

Summary of results
2001a paper: Note that this is a subset of 2001b students who also did the attitude questionnaire. Sample size 137. Students who experienced the problem-solving computer-assisted tutorial (PSCAT) had significantly higher achievement scores on the knowledge and comprehension test items than those students exposed to the lecture-internet-discussion (LID) approach. There was no statistical difference at the higher cognitive level of application, i.e. ability to apply learning to a new situation.

2001b paper: Note the same students as 2001a, but without the attitude data. Sample size 159. Knowledge and comprehension were significantly improved for those using the PSCAT; application was not improved.

2002a paper: Different sample. Knowledge was significantly improved by PSCAT; comprehension and application were not improved.

Conclusions
2001a: The results provide some experimental evidence to support the findings of previous studies that CAI generates encouraging outcomes on student achievement and attitude; in particular the ability to teach factual content. However, there was no difference between treatment and control for the higher level achievement of application of learning to new situations. In the 2001a, 2001b and 2002a papers, the author recommends that the PSCAI should be broadly developed and widely used in science classrooms.

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

The main handicap is the limited number and variety of items in the instrument and the intended generalisability.

Weight of evidence B (appropriateness of research design and analysis)
Medium-high

Reasonable sample size (4 classes) but no sampling frame/method; 'random' allocation of classes not that meaningful; very careful control of extraneous variables; good checks on reliability of data collection instrument; solid checks for data analysis, especially checking for conditions for ANCOVA but no effects sizes.

Weight of evidence C (relevance of focus of study to review)
Medium-high

Students work singly on simulations; understanding is major discrete independent variable; instrument reasonably appropriate for measuring achievement, but not for higher cognitive levels; broad range of understanding measured; work with 'private research office' not so representative of classroom teaching.

Weight of evidence D (overall weight of evidence)
Medium-high

Chang CY (2003) Teaching earth sciences: should we implement teacher-directed or student-controlled CAI in the secondary classroom? *International Journal of Science Education* 25: 427-438

Country of study
Taiwan

Details of researchers
Not stated

Name of programme (if applicable)
Assumed to be university staff member and teacher

Age of learners
11–16: 10th graders in Taiwan

Type of study
Researcher-manipulated evaluation: Controlled trial: six sets of about 40 students each were randomly assigned to either the experimental or control group. As fewer than eight sets were randomly allocated, this was not a full RCT.

Aims of study
'To develop a multimedia computer-assisted instructional tutorial in science and investigate the comparative effects of teacher-centred versus student-centred approaches on the science achievement of tenth-grade students in Taiwan.'

Summary of study design, including details of sample
Pre-test/post-test (over one week), control group experimental design, for two different treatments, the teacher-directed computer-assisted instruction method and the student-controlled method. Controlled for same teacher and same school; similar ability (implicit); similar proportions by gender (65/119 females in teacher-directed treatment, 60/113 females in student-controlled treatment); same science content; equal amount of instruction time and access to resources.

Sample = 232 students (125 females, 107 males). Teacher-directed CAI method 119 students (65 females, 54 males). Student-controlled CAI method 113 students (60 females, 53 males).

Methods used to collect data
Curriculum-based assessment: multiple-choice. 'Earth Science Achievement Test' (ESAT)

Data-collection instruments, including details of checks on reliability and validity
Through a self-constructed written 'Earth Science Achievement Test' (ESAT) with 30 multiple-choice items. 5 items cover Knowledge, 18 Comprehension and 7 Application (see also Chang 2000 and Chang 2001a).

Checks on reliability: Allocation of items to Knowledge, Comprehension and Application by six experts had high agreement (Cohen's kappa 0.74–0.84). Reliability coefficients (KR-20) of item-allocation in these 3 sub-areas based on previous data from similar sample was high (0.50 – 0.75).

Checks on validity: The instrument (ESAT) has been used (and thus validated) in several previous studies by the same author. Peer validity (six experts) was used for content validity of the items with respect to the CAI topic.

Methods used to analyse data, including details of checks on reliability and validity
Statistical analyses: MANCOVA on achievement with pre-test measures as covariates. Wilks' lambda used for testing significance of differences between treatment groups for adjusted post-treatment results (significance level 0.05).

ANCOVA on post-test achievement with pre-test as covariate on sub-level achievement (Knowledge, Comprehension, Application) to determine significance of differences between the treatment groups. Effect sizes are reported, with f = 0.1 as small, f = 0.25 as medium and f = 0.4 as large effect size.

Checks on reliability: Use of accepted statistical methods. Bonferroni method used to suppress Type I errors, i.e. significance of 0.5/3 = 0.017 as minimum accepted.

Checks on validity: ANCOVA standard and appropriate statistical test. Used the Kolmogorov-Smirnof test to check for normal distribution of test results for both groups. Used Bartlett-Box F test for homogeneity of variances among groups. Assumption of homogeneity of regression of the covariate and the dependent variable was also examined and found to be tenable. Effect size also analysed.

Summary of results
The teacher-directed CAI method was more effective that the student-controlled CAI method in improving students' achievement, particularly achievement at the Knowledge and Application levels. For the Comprehension level, there was no difference between both CAI methods.

Conclusions
'It was found that the teacher-directed CAI was more effective in improving students' achievement than student-controlled CAI, given the same learning contents and overall learning time. Moreover this study supported the notion that TDCAI used in this study significantly enhanced students' lower and higher levels in cognitive domains when compared with the SCCAI teaching method.'

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

The number of items for 'knowledge' achievement (5) and for 'application' achievement (7) are too few to be confident about the scores in these areas.

The multiple-choice test instrument is not optimal to measure comprehension and application.

Weight of evidence B (appropriateness of research design and analysis)
Medium-high

Reasonable sample size, but no sampling method, and limited generalisability; random allocation of classes, not students: attempt to control teacher, school, gender but implicit; data collection: use of a tried instrument with high KR-20 coefficients increases reliability and high agreement peer-validation for content validity; data analysis: good statistical methods including effect sizes have high reliability/validity.

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

Students work singly on PCs, or look at demonstration; understanding one of two dependent variables; multiple-choice not ideal method of testing understanding; broad range of science understanding; one intervention (student-controlled) not classroom based.

Weight of evidence D (overall weight of evidence)
Medium-high

Diehl CL (2000) *Reasoners' Workbench' Program supports students' individual and collaborative argumentation*. Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, New Orleans, LA, April 28-May 1

Country of study
USA

Details of researchers
Implicit that it is a graduate student

Name of programme (if applicable)
'Convince Me' software developed by the ECHO Educational Program

Age of learners
11–16 years: 9th graders

Type of study
Evaluation: researcher-manipulated

Aims of study
To evaluate 'the effectiveness of a computer programme, Convince Me, in supporting students' individual and collaborative argumentation'

Summary of study design, including details of sample
Students in five classes in one urban school were allocated to one of the following conditions: (1) students worked individually without feedback from the simulation model (individual, no feedback); (2) students worked individually with feedback from the simulation model (individual, feedback); (3) students worked in pairs without feedback from the simulation model (pair, no feedback); (4) students worked in pairs with feedback from the simulation model (pair, feedback); (5) control completed the same curriculum but did not use the Convince Me programme.

The curriculum was a four-week unit on waste management.

Two teachers participated in the study and, to make the teaching as similar as possible across groups, each teacher taught half the students in each of the four comparison groups. Each teacher taught one class with and one without the model's feedback.

Half the students in each class worked individually and half in pairs (N = 102).

Methods used to collect data
Curriculum-based assessment: scientific argumentation knowledge test with free responses

Data-collection instruments, including details of checks on reliability and validity
Multiple-choice five-point scale for scientific reasoning. Free answers for scientific argumentation. (Not stated how the five-point scale (strongly disagree to strongly agree) choices were converted to a score for analysis).

Checks on reliability: Pre- and post-test: Scientific reasoning test was adapted from measures on three previously published science epistemological surveys. Post-test: For the argumentation test, the free response answers were coded by two trained researchers and agreement for conflicting. Presumably computer logs were reliable.

Checks on validity: Pre- and post-test: Scientific reasoning test was adapted from measures on three previously published science epistemological surveys. No piloting of the questionnaires to this study. No details about the validity checks for the argumentation test.

Note that, of the 20 questions asked covering four categories, one of these categories was about 'relating science to real world applications'. Thus data from reasoning and belief was combined to give one overall score.

Methods used to analyse data, including details of checks on reliability and validity
Author reports that she used regression analyses. However, regression analysis would not be appropriate in the case where only two data points (pre- and post-) are collected for each experimental condition. When giving the results of the analysis, the author quotes F values (not 'r' values) so appears to have done an ANOVA test. No details of reliability are given, but ANOVA is the appropriate test. No details of validity are given, but appropriate test is used.

Summary of results
Scientific reasoning: (1) No significant differences between any of the groups for scientific reasoning in the pre-test. Total scores out of 100 are given: range from about 70 (pair, feedback) to about 72 (individual, no feedback) and about 68 for the control. (2) There was a significant difference among groups on the post-test. Estimated from Figure 3 in the report: Individual, with no feedback gave highest score about 75; then pair, no feedback about 73; then individual, with feedback 70; then pair, with feedback about 67; and control about 67. (3) Analysis of pair-wise comparisons showed that significant differences lie between the scores of the control and no-feedback groups. (4) 'The mean score ... in the Control and for Pair/Feedback groups actually decreased slightly while the mean score for the other three groups increased. A detailed analysis shows that the students in the Pair/Feedback condition did not develop less normative ideas of scientific reasoning, rather they often decreased their level of certainty...' i.e. shifted from 'strongly agree' to 'somewhat agree'. No definition or examples are given as to what the normative response would be or how it would be derived for each test question. (5) 'The four experimental groups all exhibited more normative scientific reasoning skills on the post-test questionnaire; however the two No-Feedback groups scored significantly higher than the two Feedback groups.' It is not clear why only two means are quoted when four groups were compared. Also, if this relates to the total scores data given in Figure 3, then four did *not* all exhibit more 'normative' scientific reasoning skills. There was little change for two conditions and one had a reduced total. Only the Individual/No feedback showed a noticeable increase in total score. (6) '...the only significant difference that appears in the response of the No-Feedback versus Feedback group was actually for the questions in the category of defining scientific reasoning.

Argument revisions: (1) 'Students in the No-Feedback groups were more likely to add new information to their arguments – in particular, explanatory evidence – than were students in the Feedback groups.' (2) 'Students working together were also more likely to make extrinsic changes to their argument than students working individually. (3) 'Many more students in the Feedback groups revised their existing arguments structure (72%) than did students in the No-Feedback groups (23%).'

Argument coherence ('Model's fit'): This aimed to reflect how well the students' argument structure seems to match their beliefs'. ['Beliefs' + how much the student thought their argument was plausible] used as a measure of coherence. 'The higher the overall correlation, the more the simulation agrees with the students' belief ratings – based on their argument.' (1) 'There was no significant difference among the four groups' model fit values on the initial arguments.' (2) 'However given the opportunity to reflect on and revise their arguments, the model's fit for the Feedback groups was significantly higher than that for the No Feedback groups. (3) There was very little change in the correlation between students' belief ratings and simulation ratings for the No Feedback groups after their argument revision. (4) An increase in model's fit in a student's revised argument was significantly correlated with intrinsic changes to argument structure.' 5) Students working in pairs with feedback from the simulation model had the greatest improvement in argument coherence as measured by the belief correlation value represented by the model's fit.'

Scientific argumentation: Scored for correctness and by the source of knowledge used to answer the question (a) structure, (b) content and (c) real life (not relevant for this review). (1) 'Means scores for correctness were higher for both students working with feedback from the simulation model and students working in pairs (even without model feedback). (2) 'The source of the knowledge used to answer the questions varied among the four groups with feedback from the simulation model resulting in students reflecting more on the structure of the argument in their responses.' (3) 'Students working alone without feedback were more likely to respond with real-life explanations than any other group of students.

Conclusions
Conclusions are rather general and are under the heading 'Educational Importance': 'The project described in this paper supports scientific reasoning in the classroom with computer-mediated instruction. ... Convince me aids students in generating and analysing arguments, providing feedback from a general computational model.' (1) 'Results show that in attempting to 'convince' Convince Me, students who receive feedback from the simulation model are encouraged to reflect on the structure of their arguments and their reasoning strategies. (2) 'Convince Me also appears to support reflection on argument construction and evaluation for pairs of students working together, even in the absence of feedback from the simulation model.'

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

While the design for the four conditions is appropriate for answering the study aims, the details give cause for concern. The most serious for this review is the inclusion of views on the place of science in 'real-life' applications as a measure of scientific reasoning when the other measures relate to scientific method/approach.

Using a five-point scale for student responses including normative and certainty of beliefs together, is confusing when the scoring system is not explained.

Reporting is particularly weak: no details of the students, of group sizes or of the control group, of how the reasoning skills questions were scored, means and standard deviations are not given, statistical test stated as regression when ANOVA was carried out; totals given for Figure 4 in the report when % would have been appropriate, etc.

The conclusions are vague and are not related to the findings in sufficient detail to say that the data supports the conclusions drawn, and apparent conflicts are not addressed. Data is not presented in enough detail to confirm the inferences made.

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

Large sample size (5 classes) but no sampling frame; no aspect of random control; controlled for teacher but no other extraneous variables; some checks for reliability of data collection and data analysis.

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

Study compared students working in pairs and independently; understanding of scientific reasoning wrapped up with certainty and attitude/beliefs; measures mildly appropriate for testing understanding of science ideas; reports reasonable breadth of understanding of scientific reasoning; situation representative of learners in classrooms.

Weight of evidence D (overall weight of evidence)
Medium-low

1. Dimitrov DM, McGee S, Howard BC (2002) Changes in students' science ability produced by multimedia learning environments: application of the linear logistic model for change. *School Science and Mathematics* 102: 15-24

2. McGee S, Howard BC, Dimitrov DM, Hong NS, Shia R (2002) *Addressing the complexities of evaluating interdisciplinary multimedia learning environments*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April 1-5

Country of study
USA

Details of researchers
Assumed to be university staff members

Name of programme (if applicable)
Astronomy Village: Investigating the Solar System

Age of learners
11–16: 5th grade (10–11); sixth grade (11–12; seventh grade (12–13) and eight grade 13–14 students

Type of study
Researcher-manipulated evaluation

Aims of study
'...to measure changes in students' scientific proficiency, using the LLMC, in the summative evaluation of 'Astronomy Village: Investigating the Solar System.' (LLMC stands for 'linear logistic model for change'.

Summary of study design, including details of sample
The study aimed to compare access to image analysis activities (treatment group) with no access to image activities (alternative treatment group) with the control groups. The treatment students covered only the topics related to either Mission to Pluto or Search for Life projects, which were an equivalent four weeks of all the topics in Astronomy. They participated in both the content-related activities and the inquiry oriented image analysis activities. The alternative treatment covered content-related activities which covered all of the topics in Astronomy Village over four-week period.

837 students: 590 in the Astronomy Village groups, 117 in the alternative treatment group and 130 in the control group.

12 teachers were recruited through application process and selected on certain criteria for the treatment group; 7 taught Mission to Pluto project, 5 teachers the Search for Life. Five teachers were recruited for the alternative groups. The treatment and alternative teachers were then asked to recruit teachers for their control.

Methods used to collect data
Curriculum-based assessment

Data-collection instruments, including details of checks on reliability and validity
The CET researchers developed an assessment instrument that was attuned to the topics and performance expectations of Astronomy Village: this consisted of two test scales of content understanding (22 items) and problem-solving (40 items). The instrument was a machine-readable multiple-choice format.

Checks on reliability: inquiry-related performance as the study's objectives. The researchers contracted item writers to develop assessment items. However, there is no mention of independent validation. The Cronbach's reliability coefficient was 0.08 for content understanding items and 0.97 for problem-solving items.

Checks on validity: Researchers identified the key complex content ideas presented in each of the nine investigations.

No details of establishing validity of the assessment tests given.

Methods used to analyse data, including details of checks on reliability and validity
Statistical: the linear logistic model for change (LLMC).

Checks on reliability: Use of the LLMC statistical approach.

Checks on validity: Use of the LLMC analysis, which separates any changes due to natural trends across time points (e.g. maturation effects). Results of the analysis are expressed on a logits scale.

Summary of results
No statistically significant trend effect for maturation/cognitive development over the four-week evaluation period.

There were significant effects for content understanding and for problem-solving for all three treatments. For content understanding, the greatest change was with the Search for Life), then the alternative treatment followed by the Mission to Pluto. A similar relationship was shown for problem-solving: Search for Life, then the alternative treatment, and thirdly Mission to Pluto.

'The comparison of treatment effects across scales shows that the Search for Life group and the alternative group gained more in content understanding than in problem-solving. The opposite was true for the Mission to Pluto group. The students in that group did slightly better in problem-solving than in content understanding.

Conclusions
'The results from this study indicate that the material developed for the Astronomy Village project can be used effectively to promote interdisciplinary understanding and problem solving in planetary science within a relatively short time.'

The Search for Life students studied half as much content as the alternative group but they improved slightly more than the alternative group on the content understanding test. In some cases, the Search for Life students used the image analysis and were better able to transfer that to problem-solving Mission to Pluto problems in the test than were alternative treatment students. This supports the view that students should study fewer topics but in more depth using an enquiry based approach.

The Mission to Pluto treatment offered fewer activities that would transfer to Search for Life problem-solving. In this treatment, students spent more time learning how to interpret remote sensing of planetary features and how to infer underlying processes.

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

Large sample size and some care in choosing sample but taken from within sample of volunteering teachers. No details on how the students work with the IT – individually, in groups, as a class together? This was probably because it varies across schools. No information on what the No Treatment group experienced. Reliability of test instrument checked. Less information on validity. Strong on data analysis. Conclusions in line with results.

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

Large sample size but researchers had limited control of sample (subset from volunteering teachers/schools); no aspect of random control; some control of extraneous variables (gender, socio-demographic and grades); standard checks on reliability of data collection but less on validity; used valid statistical test to enhance the reliability of the data analysis.

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

No details of how the students worked with the IT; understanding is sole focus; measures appropriate for understanding of science ideas; reports broad range of science ideas; no details of classroom situation.

Weight of evidence D (overall weight of evidence)
Medium

Huffman D, Goldberg F, Michlin M (2003) Using computers to create constructivist learning environments: impact on pedagogy and achievement. *Journal of Computers in Mathematics and Science Teaching* 22: 151-168

Country of study
USA

Details of researchers
Study written by university staff, but presumably data collected by the teachers in the schools as the evaluation was carried out across 23 schools.

Name of programme (if applicable)
Constructing Physics Understanding (CPU) project

Age of learners
No specific age given but were high school students. Study was done in 23 schools so ages may have varied from school to school.

Type of study
Evaluation: researcher-manipulated

Aims of study
To compare the extent to which teachers using the CPU (Constructing Physics Understanding) project, and teachers not using the project materials could create a constructivist learning environment.

Summary of study design, including details of sample
The CPU programme was used to teach a physics unit. In one category, classes were taught by an experienced CPU teacher; in another, classes were taught by a newly trained CPU teacher; and comparison/control classes were taught by a teacher using non-CPU materials.

A pre-post achievement test was used to investigate any differences in students' understanding of force and motion.

Sample: 23 high schools; 9 CPU lead classes with 172 pupils taught by 4 different teachers; 8 beginning CPU classes with 116 students and 4 different teachers; 6 comparison classes with 78 students taught by 5 different teachers.

Methods used to collect data
Curriculum-based assessment

Data-collection instruments, including details of checks on reliability and validity
Force concept inventory (FCI) test of force and motion; a multiple-choice test

Nationally recognised test was chosen because it includes a wide range of concepts taught, has a good reliability and because there are existing national data on how students in other high school physics classes have scored on the test.

Checks on reliability: The test has a Cronbach alpha in the range of 0.70.

Checks on validity: This is a nationally recognised test.

Methods used to analyse data, including details of checks on reliability and validity
Each multiple-choice test question was given one point up to a maximum of 30 points. Chi-squared at the start to look for any demographic differences between the groups. No differences were found.

Analysis of variance and follow up 't' tests for physics understanding. Cohen's effect size analysis for physics understanding.

Hake Plot analysis of percentage gain versus percentage pre-test score was used. (This takes into account that students with lower pre-test scores have more opportunity to gain than students with higher pre-test scores.)

Checks on reliability: Reliability is built into the tests used.

Checks on validity: Chi-squared, ANOVA and 't' test are the standard test to use for this type of analysis.

Summary of results
Statistically significant differences between the groups on the pre-test, post-test and gain scores.

On the post-test, the students in the lead teachers' classes had the highest scores, the students in the beginning CPU classes had the next highest scores, and the students in the comparison classes had the lowest scores. A similar pattern was found for gain scores.

Cohen's effect sizes indicated that there were large to medium differences between beginning CPU students and comparison students, and between lead CPU students and comparison students.

Compared against data from national norms, the lead CPU classes fall just below the line of other highly interactive classes on the Hake plot; the beginning CPU classes fall somewhere between other highly interactive classes and lecture-dominated classes; and the comparison classes fall below the traditional line.

However, CPU use was hindered by technical challenges of using computers, and gaining access to computers to use. Teachers found the new pedagogy associated with CPU difficult.

Conclusions
'…the CPU project significantly improved students understanding of physics concepts. When compared to traditional classes, students in CPU classes make significantly higher gains in their understanding of concepts. Relative to national norms, however the classes of beginning CPU teachers fell somewhere between highly interactive classes and lecture-dominated classes. The classes of lead CPU teachers had scores approaching other highly interactive classes. These results suggest that CPU had a positive impact on students' understanding of physics concepts, but was not able to improve students' achievement as much as other interactive approaches.'

Also computers can be used to help teachers create a more constructivist learning environment, although this is challenging for them.

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

This was a carefully designed study with a cross-section of schools and good sample size. However, hardly any detail is given of the sample. It was established statistically that the student characteristics for the two groups of students of particular interest (two levels of treatment) were not different and the control (no treatment) group was matched for teacher and demographic characteristics of the students.

A core part of the study was the level of CPU (IT) experience of the teachers, which could otherwise have been an extraneous variable.

Appropriate tests were carried out for reliability and validity of data collection and analysis. Data was also compared with national norms.

Conclusions were supported.

Weight of evidence B (appropriateness of research design and analysis)
Medium-high

Large sample but no sampling frame; no details of how the students or classes were chosen; did control for type of teacher and checked that student demographic characteristics in the three samples were the same; solid checks for reliability and validity of data collection and of analyses.

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

Students using CPU work in small groups; understanding is a major discrete element of intervention; achievement test is highly appropriate for testing directly the understanding of forces and motion; wide range of concepts tested; situation highly representative of students in the classroom.

Weight of evidence D (overall weight of evidence)
Medium-high

Huppert J, Lomask SM, Lazarowitz R (2002) Computer simulations in the high school: students' cognitive stages, science process skills and academic achievement in microbiology. *International Journal of Science Education* 24: 803-821

Country of study
Israel

Details of researchers
One based at US university, one at an Israeli university and one at an Israeli technion

Name of programme (if applicable)
The Growth Curve of Microorganisms

Age of learners
Tenth graders; 15–16 years old

Type of study
Evaluation: researcher-manipulated

Aims of study
To implement a computer simulation program 'The Growth Curve of Microorganisms' in tenth-grade biology classrooms and investigate students' academic achievement and the mastery of science process skills, in relation to their cognitive operational stage and their gender.

Summary of study design, including details of sample
Five tenth-grade classes in Israel were divided into two groups, each with males and females.

Experimental group were taught about the growth curve of micro-organisms using a combination of classroom teaching, laboratory experiments and computer simulation experiments. The control group was taught using the traditional classroom/laboratory method.

No details were given of how many schools were involved or of how the classes were allocated to the two groups.

Sample: Experimental group of 82 (68 girls and 14 boys). Control group of 99 (80 girls and 19 boys).

Methods used to collect data
Curriculum-based assessment: pencil and paper tests for all three measures.

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

- To define the sample: Student cognitive stages were assessed using a Video-Taped Group Test. This 12-task test requires formal reasoning powers: conservation of weight and volume, control of variables, proportional, probabilistic, combinatorial and correlation reasoning.

- To measure aspects of the sample as findings of the study: (1) Knowledge – pre-test was 40 multiple-choice questions; 30 on general knowledge in biology and 10 about previous knowledge on the topic of the growth rate of micro-organisms. The post-test was on the growth rate of micro-organisms mastered during the learning process. There were 15 multiple choice and five essay questions. All related to the topic of micro-organisms, structure, size and various factors affecting population growth. (2) Science process skills were measured using the Biology Test of Science Processes. This instrument contains nine sub-scales with a total of 48 items/questions.

- Reliability: (1) Knowledge tests: pre-test, using the Spearman-Brown formula gave a reliability value of 0.89. Also student answers were analysed and compared with a key answer prepared by the investigators. Post-test reliability was established using Cronbach test. (2) Cognitive stages reliability was established. (3) Science process skills reliability was established by Lazarowitz and Huppert (1993) for the nine sub-scales.

- Validity: (1) Knowledge tests: pre-test validity established by five teachers, who reached an inter-judgement agreement of 85%. Post-test validity was established using teachers, who gave a 90% inter-judgement agreement. (2) Cognitive stages: validity for tenth-graders was established, using five science educators with an inter-judgement agreement of 91%. (3) Science process skills: validity was established using five high school teachers.

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

- T-tests for general knowledge in pre-test established that there was no significant difference between the groups.

- To categorise the cognitive stages for the science process skills analysis subjects could score 0–2 points for each task and 0–24 points for completing the test. Students were then divided into three categories: 0–8 points = concrete thinkers; 9–16 points = transitional thinkers and 17–24 points = formal reasoners.

- Two-way analysis of variance and Sheffe test for post-test science achievement and for science process skills. Multiple analysis of variance was used to analyse the nine subscales for the science process skills. Effect sizes were calculated.

- Reliability is built into the statistical tests.

- Validity: appropriate tests are used to ensure validity.

Summary of results

Pre-test

- Students' mean scores in the pre-test of general knowledge in biology and of previous knowledge in the population growth rate of micro-organisms indicated no significant differences between the experimental and control groups.

Subject knowledge

- Students of the experimental group in the concrete and transitional operational stages achieved significantly higher mean scores than their counterparts in the control group. No significant differences on the mean scores between the experimental and control formal operational students.

- Within each study group, the following picture emerged: (a) The higher the cognitive operational stage, the higher was the students' academic achievement. (b) No significant differences on the mean scores were found between the students in the concrete and transitional operational stages within each group. (c) Within the experimental group and within the control group, significant differences in the mean scores were found between the concrete scores and formal scores. (d) Within the control group (but not the experimental), there was a significant difference between the scores of the transitional and formal reasoners.

Science process skills

- There was only one significant difference with gender and cognitive stage: girls of the transitional stage in the experimental group achieved significantly higher mean scores than their counterparts in the control group.

- Science process skills analysed by nine sub-scales: (a) The concrete experimental group have significantly higher mean scores for the skills of measurement, graph communication, interpreting data than the control group. (b) The experimental group in the transitional cognitive stage have significantly higher scores for the sub-scale skills of measurement, classification, graph communication than the control group. (c) Formal cognitive stage – there were no differences between the experimental and control groups for all nine skills.

Conclusions

Between experimental and control:

- For scientific knowledge, both concrete and transitional reasoners did better with the simulations than without. There was no difference for formal (more advanced) reasoners between the experimental and control group.

Within experimental and control:

- When achievement was analysed within groups, it was found that for both groups the transitional and formal groups did better than the concrete groups.

- For process skills, experimental groups did better than the controls for concrete and transitional groups in respect of the lower end skills of measurement, graph communication and interpreting data.

- Formal thinkers showed no differences for any of the nine skills between those using the simulations and those not.

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

Medium

The strength of this study is the use of established knowledge and science process skills tests which were checked for reliability and validity. Three appropriate tests were used for the data analysis and details of the results provided in the text and tables. More detail could have been given about the sample, its recruitment and allocation to experimental and control conditions. However, there were six out of 21 cases where the level of significance was misinterpreted and stated to give a significant result.

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

Medium

Reasonable sample but no sampling frame; no aspect of random control; checked for pre-test knowledge and controlled for gender; solid checks for reliability and validity of data collection; used four appropriate methods of data analysis but did not use standard probability threshold of P = 0.05.

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

Medium-high

No details of whether students worked with the software singly or in groups; understanding (of science content and of science process) are sole focus of the study; measures are highly appropriate; reports good breadth of knowledge and process skills; students work in classrooms.

Weight of evidence D (overall weight of evidence)

Medium

Miglino O, Rubinacci F, Pagliarini L, Lund HH (2004) Using artificial life to teach evolutionary biology. *Cognitive Processing* 5: 123-129

Country of study
Italy

Details of researchers
One researcher at an Italian university, one at a Danish university and one at a national institute for cognitive science and technologies in Italy

Name of programme (if applicable)
Suite of artificial life software: FACE-IT (genetics software package); TOYBOT (simulation of interesting mobile robot behaviours); SURVIVE! (a software package containing a micro-world of artificial organisms and food)

Control software was 'Charles Darwin'.

Age of learners
14–15 years old

Type of study
Researcher-manipulated evaluation

Aims of study
To pilot the use of artificial life software in an educational setting and obtain an initial measure of the effectiveness of the performance of learners using 'artificial life' tools (experimental group) compared with a group using traditional multimedia hypertext (control).

Summary of study design, including details of sample
The experimental and control groups each consisted of 22 high school students aged between 14–15 years.

No details of school, ability or how the students were allocated to the groups. One teacher for the two groups.

Each group was given a standard lesson on evolutionary biology by their teacher. Then the control group used a commercial hypertext while the experimental group used the artificial life software.

Methods used to collect data
Curriculum-based assessment

Data-collection instruments, including details of checks on reliability and validity
- Multiple-choice questionnaire of 14 questions on basic notions of Darwinian theory. Each with four possible answers offered, only one of which was correct. Whole questionnaire is shown in Appendix 1 of the report.

- No checks on reliability or validity.

Methods used to analyse data, including details of checks on reliability and validity
- Statistical analysis of differences in performance scores before and after a standard lesson within groups. Then statistical analysis of differences in performance scores after the software lesson both within and between groups (hypertext and artificial life).

- Analysis of variance. Standard and appropriate test for the data analyses.

- Checks on reliability and validity: test is standard and appropriate one.

Summary of results
- Both groups started with the same level of knowledge.

- Both groups improved their performance after the teacher's lesson.

- Following the software sessions, the students achieved a further increase in their performance.

- Artificial life users scored significantly higher marks then hypertext users.

Conclusions
- Complementing traditional teaching with other educational tools (in this case hypertext and artificial life software) can have a significant effect on learning performance.

- Students who used artificial life software (virtual experiments) achieved greater improvements in performance than those who used hypertext.

- However, the authors point out that they could not rule out the possibility that the observed effect was 'simply due to the extra time spent on the subject'. (No detail was given about the extent of this extra time.) They suggest that, in future research, a control group takes additional traditional lessons for a time equivalent to the duration of the software session.

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

This pilot study set out to compare two types of software: one involving interactive, virtual experiments (artificial life) and the other the more passive, multimedia hypertext. This was accomplished. The additional control (traditionally taught lesson time equivalent to the time spent on software) suggested by the authors would provide further worthwhile information but would not be crucial to this study. However, the sample was small and there was little reported attempt to control the sample for gender, academic ability or socio-economic factors. It was not reported how long the students spent with the software.

There are concerns about the equivalence, or not, of the content of the two types of software and about the extent of the relationship (validity) of the test questions to the hypertext/control material. The authors needed to discuss this.

Weight of evidence B (appropriateness of research design and analysis)
Medium-low

Small sample size (44 students); no details of how allocated to groups; controlled for teacher and school but no details of gender or achievement level; did not check for the reliability or validity of the questions in the test in relation to the content of the two types of software; data analysis was standard and appropriate.

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

No details of how (singly, pairs, groups) the students worked on the simulations; focus of the intervention is the sole variable; measures appropriate for testing understanding but not sufficiently closely matched to the content of the control software; breadth of questions appropriate for testing understanding of biological evolution; no details given of where/when/how long students worked on the software.

Weight of evidence D (overall weight of evidence)
Medium-low

The results of this systematic review are available in three formats:

SUMMARY	Explains the purpose of the review and the main messages from the research evidence
TECHNICAL REPORT	**Includes the background, main findings, and full technical details of the review**
DATABASES	Access to codings describing each research study included in the review

These can be downloaded or accessed at **http://eppi.ioe.ac.uk**

First produced in 2006 by:

Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre)
Social Science Research Unit
Institute of Education, University of London
18 Woburn Square
London WC1H 0NR

Tel: +44 (0)20 7612 6367

http://eppi.ioe.ac.uk/
http://www.ioe.ac.uk/ssru/

The **Evidence for Policy and Practice Information and Co-ordinating Centre** (EPPI-Centre) is part of the Social Science Research Unit (SSRU), Institute of Education, University of London.

The EPPI-Centre was established in 1993 to address the need for a systematic approach to the organisation and review of evidence-based work on social interventions. The work and publications of the Centre engage health and education policy makers, practitioners and service users in discussions about how researchers can make their work more relevant and how to use research findings.

Founded in 1990, the Social Science Research Unit (SSRU) is based at the Institute of Education, University of London. Our mission is to engage in and otherwise promote rigorous, ethical and participative social research as well as to support evidence-informed public policy and practice across a range of domains including education, health and welfare, guided by a concern for human rights, social justice and the development of human potential.

This document is available in a range of accessible formats including large print. Please contact the Institute of Education for assistance:

telephone: +44 (0)20 7947 9556 email: info@ioe.ac.uk