Debating the Evidence

A Futurelab prototype research report

by Keri Facer, Mary Ulicsak (Futurelab); Paul Howard-Jones (University of Bristol)

August 2005

CONTENTS

1. Executive summary
2. Overview of Debating the Evidence development process
3. The development of the research questions
4. Summary of the final 'learning environment' created
5. Evaluation of Debating the Evidence
6. Analysis
7. General findings
References

1. EXECUTIVE SUMMARY

Debating the Evidence is designed to raise awareness of risk and uncertainty in scientific reasoning and support students' collaboration in engaging with these issues.

The learning environment comprises three stages: 1) in pairs, students aged 11 to 14 first complete training exercises which involve analysing data with covariance between cause and effect. To assist the students the software provides feedback about the strategy they employed based on their inputs; 2) then, once mastered, students analyse findings from a data set in which there is limited covariance between factors, that is, there is an imperfect relationship between cause and effect; 3) finally, the students have to draw conclusions and present their findings and recommendations. Awareness of risk and uncertainty is raised in the second activity, which is closer to 'real' scenarios such as those dealing with evidence related to genetically modified food or vaccination. In this second activity, the students have to reason why discrepancies occur and how much evidence is sufficient, given the cost and consequences of delaying recommendations. To support students' collaboration, the software requires that **both** students participate, as individuals and by providing an agreed hypothesis for causes and outcome. The software provides feedback on the apparent amount of cooperation between students' when providing 'agreed' responses.

The software is accessed via a webpage, however, the inputs and feedback are logged on a server and can be accessed by the students and teachers at a later time. Ideally, the personal computer is set up with two mice so that students have an independent means of entering a decision.

After consultation with teachers and a usability study with gifted and talented students, Debating the Evidence was trialled with a Year 8 class. The study focused on the impact of the software rather than the creation of a learning environment in which the software would be embedded with additional teaching and material resources. This report summarises the process and findings.

1.1 Key innovations of Debating the Evidence

Debating the Evidence was designed to be innovative in three key areas, as summarised below:

1.1.1 Engagement with uncertainty

It was intended that this would be achieved via collaborative prediction-making based on evidence that was partially inconsistent. It was intended that such an experience would improve students' awareness of the importance and limitations of scientific evidence.

1.1.2 Dual responsibility

The software was designed for pairs, but theories and predictions were entered first individually and then as a response agreed by the pair. Ideally, each student would have their own mouse for inputting their individual response. This was intended to support collaboration through committing individuals to form personal as well as negotiated theories. It also facilitated analysis of inter- and intra-individual strategies by the system and the production of formative feedback.

1.1.3 Automated formative feedback

The software provided students with prediction feedback and also formative feedback regarding the problem-solving strategies demonstrated and the extent of their peer co-operation, that is, how they appeared to choose the joint response. Formative feedback was carried out by the system via a dynamic analysis of the students' responses. This was intended to improve students' consideration of scientific evidence and the extent to which they collaborated on the problem-solving task. Analysis of the students' behaviour used the

relationship between their theories and their predictions, and the sequence in which these occurred, to characterise their thinking strategies and provide appropriate advice to the students about how these could be improved.

1.2 Key learning findings and recommendations

The process of developing and trialling the software led to a better understanding of how students comprehend uncertainty and the impact of the feedback. It also highlighted how such software should be designed and areas for future investigation. The key findings are listed below.

1.2.1 For the educational research community

- Interaction with a computer simulation of a scenario where there is an imperfect cause-effect relationship appears to improve students' ability to sceptically examine how evidence is used.
- Rapid feedback from their prediction outcomes engages and supports students in revising their theories.
- Requiring individuals to enter their own responses, as well as their agreed ones, allowed the system to monitor inter- and intra-individual performance as well as allowing individual explicit expression. The pupils appeared motivated to do this, and may even have been motivated **by** this in their collaboration, but were **not** motivated to do this as soon as the decisions became routine.
- In the absence of teacher-intervention, instances were recorded where students improved their thinking strategies following relevant automatic formative feedback from a dynamic computer-based analysis of their behaviour. However, further research is needed to investigate the role that the automated feedback played in this improvement.
- Although completing the training facilitated the dynamic analysis of the students' strategies and thus the provision of formative feedback, the students' motivation dropped during training as soon as they were confident that they had identified the two causes. This contrasted with the mission, where the element of unpredictability, even when causes had been correctly identified, maintained the students' interest in sharing opportunities to make further predictions and increase the evidence base.

1.2.2 For teachers, advisors and head teachers

- Students appear to find interactive encounters with simulated scientific problems involving uncertain cause-effect relationships interesting and challenging, and this project has produced some evidence that this interaction improves their critical consideration of how evidence is used.
- The heterogeneity of outcomes within the class and the diverse experiences that different pupils had with the same software also suggest that the experience could be a useful precursor to classroom "science in society" debates about the importance and limitations of scientific evidence.
- Debating the Evidence supports the 21st Century Science curriculum being introduced in September 2006, as it focuses on the understanding that can be applied rather than the acquisition of facts.
- The *training* was of most benefit to students who had most difficulty hypothesis testing and prediction making. While the *mission* was of most benefit to those students who already had some understanding, albeit implicit, that inconsistent evidence weakens confidence in findings.

1.2.3 Policy makers and industry

- Students often demonstrated an unreasoned mistrust of authority, echoing some of the popular rejections of government and medical advice on recent issues such as the measles,

mumps and rubella (MMR) vaccination. This may reflect a need for greater attention to be given in the curriculum to the critical analysis and appreciation of scientific evidence in areas of social concern.
- Students are not generally prepared in schools to make reasoned assessments of confidence based on incomplete evidence and interactive software may play a key role in helping them make reasoned judgements about situations involving uncertainty.
- Even without teacher intervention children can make gains by working collaboratively using a system that provides feedback on their performance and teamwork.

2. OVERVIEW OF DEBATING THE EVIDENCE DEVELOPMENT PROCESS

Debating the Evidence was accepted in the Call for Ideas application process in Spring 2004 as a citizenship project. However, the project was not commenced until the following autumn. The submitted goal was to "produce a simulation that engages students with collaboratively reviewing and presenting evidence, recommendations and findings, and to discuss notions of risk and uncertainty, thus supporting their understanding of science-in-society issues. In stages of increasing social and scientific complexity, we want to challenge young adults to become involved in debates involving scientific evidence and human values" (Howard-Jones 2004).

Activities that directly influenced the pedagogical design of the software, and the final trials, thus any sessions with informants, that is, students and teachers are listed in Table 1.

Prototype Stage	Dates	Participants	Purpose	Summary of key outcomes
Plot scenario	15 Nov	Paul Howard Jones, Mary Ulicsak		• Identification of plants and mutant cats as environment for testing understanding of causality
Wire frames	Oct – Dec 2004	Richard Caddick, Pete Ferne, Clara Mortimer	Identify functionality of each screen	• Specification model for developers
Initial teacher interview	9 Nov 2004	Duncan McCalmont (teacher) and Keri Facer	Review	• Students will need scaffolding to use all the features • The task will need introducing • It may be used, but not often – issues with how it could be marked
Teacher review	7 Mar 2005	Duncan McCalmont, Sarah Richards	Review wording and how to integrate into a lesson	• Lesson plan required – how does it fit in with curriculum • Animated sequences of instructions would be useful • Pro forma for information for presentation should be provided
Usability study	18 March 2005	7 gifted and talented students: 2 Year 7 students (1 boy, 1 girl) 4 Year 8 students (2 boys, 2 girls)	Review proposed lesson plan and software	• These Year 8 and 9 students achieved the task without problems • Guessing which was the rogue plant or cat initially was counter-intuitive • That the instructions were

		1 Year 9 student (1 boy)		rarely referred to and a demonstration of use would be more appropriate • That the feedback was ignored unless flagged • That having to complete the 20 tests in the training became tedious to students once they had identified the causes and were achieving 100% predictions

Table 1: Overview of learning development process

In addition, between October and December the functionality of each screen was identified using wire frames. And in November there were discussions about the graphical representation of the software. This was reviewed by those involved within the project team rather than informants. In December the underlying architecture was finalised, that is, the algorithms for providing feedback and the storing of files to be accessed by the teacher. Again with respect to the teacher pages there was no confirmation of the suitability by anyone outside the project team.

3. THE DEVELOPMENT OF THE RESEARCH QUESTIONS

Debating the Evidence was designed to address four areas that students may find problematic: thinking scientifically about evidence, working collaboratively, interpreting feedback and responding to unpredictability. This section summarises the literature review which discusses these issues and concludes with the research questions that were identified (a full version of the literature review is available on the Futurelab website)

The literature review discusses the need for software that supports reasoning, as students often appear not to test hypotheses in a systematic manner and have considerable difficulty in understanding and applying the scientific method. They may not seek out information that could disprove their hypotheses and may easily accept causes that only partially account for the evidence. In one study 11 to 14 year-olds were shown pictorial evidence showing groups of children who enjoyed different diets and different states of health. As in other studies involving adolescents the older children had some success in using the covariation information effectively, but the younger children in this age range found the task very difficult. Inclusion errors - in which individuals wrongly base their conclusions upon a single instance of a variable covarying with the outcome - accounted for a large proportion of the mistakes made. The importance of promoting children's abilities to consider evidence, such as test outcomes, when attempting to identify cause is now emphasised in science curricula. Strategies intended to enhance such reasoning skills include the encouragement of peer collaboration.

Research into structuring collaborative environments is ongoing. Collaboration is commonly defined as any situation in which two or more people attempt to learn something together, a more precise definition is that collaboration is a coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of the problem. Recent research into collaborative learning has emphasised the importance of the socially mediated processes of conflict resolution and the negotiation of meaning, rather than simply exposure to conflict. With regard to working at computers it has been demonstrated that those children who had been encouraged to share the keyboard, help each other and discuss and compare ideas, achieved higher results in a post-test based on similar problems than those children who had worked on their own. However, some studies conclude that even if collaboration occurs, the students appear to be communicating and producing more scientifically correct concepts, this does not necessarily imply the students have better

individually internalised representations. Neither does it mean that the task is motivating and engaging.

Other studies suggest that peer-group collaboration may not lead to improvements *unless* combined with other factors such as guidelines and feedback. Unguided talk around computers is of limited educational value, instructions must be appropriate and the use of procedural knowledge and procedural skills must be integrated. A collaborative task is beneficial if it leads to talk that supports learning. Whether it is more beneficial for this talk to be a self-explanation of actions by individuals, or asking them to explain what they observed others doing, or having to explaining to a peer their approach is unknown. Regardless of format this suggests that it is important to structure a task so that communicating meaningfully about the activity is integral.

Supporting discussions relies on displaying the evidence appropriately. However, there are no clear guidelines on the ideal format for this. Multimedia simulations that provide animated automatic feedback about the effects of the learners' actions have been positively evaluated but the use of any multimedia approaches to providing feedback should probably be tempered by an awareness of the principles of cognitive load. Interestingly, cognitive load principles have also been used to support the argument that explanatory feedback in creative problem-solving is of greater help to the learner than simple corrective feedback, since it relieves some of the heavy cognitive burden of exploring a highly complex environment. The predictive feedback proposed should support students in determining their own level of confidence in their theories about cause. They found that use of feedback by students was increased at lower levels of certitude.

Finally, in many real-life situations there is an element of uncertainty. For example, there has been some difficulty, to a greater or lesser extent, in completely proving or disproving cause-effect relationships in high-profile cases such as MMR vaccinations, genetically modified food and the impact of eating beef from cattle with bovine spongiform encephalopathy (BSE). In the presence of evidence that appears inconsistent, children and adults will fall back upon behaviours that are more easily influenced by 'social reasoning' phenomona. For example, an individual may become more likely to attribute their own erroneous predictions to difficulties inherent in the problem (eg it's a difficult situation to make correct predictions), while those of their collaborator may be attributed to their personal qualities (they're just not good at it). This "fundamental attribution error" (Ross 1977) may contribute unhelpfully to the grounding required for productive collaboration and for the objectivity needed to interpret the evidence as appropriately as possible. Furthermore, in the social domain, the ideas we verbalise may not be representations of the reasoning we use to guide our behaviour and future predictions need not be in line with past ones. However, these less scientific strategies, such as pattern matching, may be more helpful in some social situations.

From the above it appeared that software to support the development of reasoning around scientific thinking and uncertainty was required. This software should support collaboration by requiring the participants to communicate and exchange information that is meaningful and integral to the task; hence the introduction of individual and joint responsibility for hypothesising and identifying outcomes in Debating the Evidence. The feedback provided by the software must be sufficient, so as not to introduce a heavy cognitive load, but contain enough information to be useful. It should also make the task more engaging. The proposal was that the medium of the feedback to be displayed could be selected by the teacher. For example, students working randomly could have a picture of random squiggles displayed from which they would have to deduce the meaning, or those that had prior belief would be shown as static against a backdrop of changing patterns. This would have been in conjunction with visual cues like the distance between their feedback illustrating the amount of agreement in joint decision making. This led to the identification of three specific research questions which could be addressed by the Debating the Evidence software:

- How does the proposed framework support the development of collaboration?

- Does the type of collaboration supported by the proposed framework transform the student's understanding of scientific thinking and uncertainty?
- What effect does the feedback, presented in different modes, have on motivating and engaging the students?

4. SUMMARY OF THE FINAL 'LEARNING ENVIRONMENT' CREATED

4.1 Description of activity

Debating the Evidence is designed for pairs of students aged between 11 and 14 working at the same internet-connected computer. The initial stage, the training, occurs in a laboratory setting. Here, students must identify causes of premature plant death in a controlled situation where there is a reliable cause-effect relationship. The teacher can set the complexity of the training and the number of blocks that the students must complete. In this level the system provides almost constant feedback in terms of the outcomes of tests that the students have made predictions about. As well as prediction feedback, formative feedback is given after each block of tests on the extent of the co-operation between the two students and also on their individual and agreed problem-solving strategies. In the implementation tested the only method for providing feedback was via text. Students do not pass the training until both participants appear to be adopting a mature scientific approach.

In the second stage the students work on the Katzville mission, where they undertake a 'field study' in less controlled conditions. Here, the task is to identify the cat or cats with rogue genes whose offspring usually become violent. Unlike the plant scenario encountered in training, the teacher can add uncertainty by setting the software to produce a number of anomalous test outcomes and select the cat(s) that are rogue. Thus, when identifying causes, the students must implicitly or explicitly agree a level of confidence about their deductions that takes account of the contradictions in their experimental results. The system does not provide any formative feedback on problem solving strategies but will produce messages when the level of co-operation is inappropriate. That is, it sends a message when students consistently diverge in their answers and do not take into account the others opinion when entering the 'agreed' response.

In the original proposal there was to be a third stage (Howard-Jones 2003). The pair of students would have to present their findings and conclusions. This had been intended as an extra incentive and opportunity to reason about the evidence they collected, as well as to transfer insights and create new ones. However, due to curriculum constraints within the school chosen for the evaluation, it was not possible to evaluate the use of the software to support classroom debate.

The key features of the Debating the Evidence activity include:

- individual and joint responses required, ideally by dual key control
- the software provided students with prediction feedback
- provision of formative feedback regarding students' problem-solving strategies and the extent of their peer cooperation.

The time to complete the activity is dependent on the scenario provided by the teacher.

4.2 Description of prototype

Debating the Evidence can be run on a standard PC attached to the internet as it is accessed via a web page. It was designed to have an additional mouse to be attached to the machine for the dual key control. The teacher can access the log files from the activities from the teachers log-in page.

After initiating the software, students were asked to enter their names at the opening screen, before the home page is reached. On this first visit to the home page, they were guided by the teacher-researcher to choose the 'training' option. When they had successfully completed training, they would be returned to this page having been prompted by the computer to choose the 'mission' option.

4.2.1 Training – simulated laboratory studies

Having chosen the training option, and after observing a brief picture of the entrance to the training laboratory, the training screen appears. Details about the background context and how to train is available by clicking on the handbook displayed on the screen.

In this clean and clinical laboratory setting, students observe tests of plants involving crossing two plants. The total number of plants is set by the teacher, but must be between three and six; they also select the number with rogue genes, which must be one or two. The total number of plants and the number containing rogue genes is displayed at the top of the screen. So, in the case where there are two suspected types, combinations that contain either or both of these die within six months. For example, if there were five plants they would be labelled A, B, C, D and E and produce cross-breeds of AD, CE etc. If only the original A plant contains the rogue gene, then AD dies but CE does not. The software chooses the causes randomly so only collaboration within, rather than between, pairs is helpful.

Before each test, the left and right students enter their ideas individually as to their current theory about which is/are the rogue plant(s), and then their agreed idea. Prompts for individuals to submit theories appear randomly at the appropriate side on the bottom of the screen. The students are then prompted to give a theory as a pair. Responses are made by clicking on the options appearing on the bottom of the screen. This is shown in Figure 1.

Figure 1: Request to pair for hypothesis, screen shows prediction evidence on whiteboard in background, test rig, and handbook with instruction

A combination then appears on the on the computer monitor shown in Figure 1, eg 'B + E'. This is apparently chosen randomly by the computer, and left and right students enter their individual predictions regarding the outcome. After then entering their agreed prediction, the development of the plant is accelerated in the RDA (Reversible Developmental Accelerator) and the outcome, Lived or Died (indicated by whether the plant shrivels up or not) is stored on the

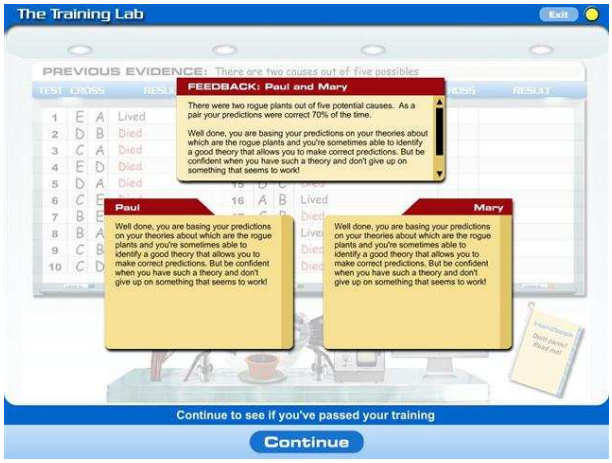
whiteboard above the test bench. This process is then repeated, with evidence supporting the students' theories and predictions accumulating on the whiteboard. The tests occur so that in both halves of the block every combination is tried once, but the order of the plant identifiers and occurrences is random.

At the end of each block, the system identifies the strategies applied by the students to solve the problem. This formative feedback is designed to support the progression from non-optimal strategies to a mature scientific deduction of cause. This feedback is automatically produced by the system through measuring theory-prediction consistency (TPC) and analysing the sequences of theories expressed and predictions made. TPC is defined as the percentage of the predictions made which would be those expected to arise from the theory held. The software was able to identify the five types of strategy shown in Table 3 below along with the message it generates:

Strategy	TPC	Definition	Feedback generated
Random	< 75%	Default (no TPC)	Think about your approach. Think about how you can work out which is a rogue plant and/or which isn't. Base your predictions on whether you think the rogue {plant|plants} {is|are} present in the pair being tested.
Pattern matching		75% or more of predictions are correct in the second half of the block.	You're basing predictions on what happened when the same pair was tested. You can get a higher prediction score by trying to identify the rogue {plant|plants} and basing your predictions on whether you think the rogue {plant|plants} {is|are} present in the pair being tested.
Prior belief	>= 75%	The results block ends with a long string (this is a tunable parameter dependent on block size) of incorrect predictions, based on an incorrect theory and no change of theory.	Well done, you are basing your predictions on your theories about which {is|are} the rogue {plant|plants}. But don't hang on too long to a theory that doesn't give good predictions - its probably wrong!
Vacillation		TPC and maintaining the (both) correct theory (theories) for 2 or more occasions but then departing from it; or TPC but changing theory (at least one theory) after one or more correct predictions.	Well done, you are basing your predictions on your theories about which {is|are} the rogue {plant|plants} and you're sometimes able to identify a good theory that allows you to make correct predictions. But be confident when you have such a theory and don't give up on something that seems to work!
Mature		The results block ends with a long string of correct predictions, based on a correct theory and no change of theory.	Well done, you are basing your predictions on your theories about which {is|are} the rogue {plant|plants} and you're able to identify a good theory that allows you to make correct predictions.

Table 2: Summary of problem-solving strategies

The feedback can be edited by the teacher. The system also identifies instances when the agreed decisions are heavily biased towards the individual responses of either the left or right student, and feedback is given to encourage a more equitable approach. The teacher can choose to set up the system so the distance between the messages depends on the amount of shared inputs, as shown in Figure 2.



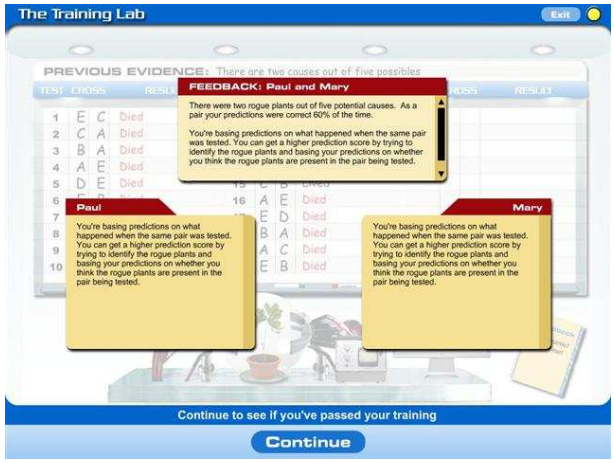


Figure 2a: Feedbacks prompts when pair appeared to make decisions jointly

b: Feedback when pair did not appear to make decisions jointly

When a mature strategy is achieved and sustained by both individuals in their last training block (they can have up to four blocks) a message appears that they have passed their training and should return to the homepage to start their mission. Otherwise they are encouraged to repeat their training.

4.2.2 The Mission – a simulated field study

Having chosen the mission option either from the introduction screen or after successfully completing training, and after a brief picture of a dirty street in Katzville, the mission screen appears. Details about the background context and how to carry out the mission are available by clicking on the handbook displayed on the screen.

The handbook explains that students are now trained as scientific investigators of crime and are being sent on a mission in a time not far in the future, when genetically modified pets are abundant. One or two of three, four, five or six original GM cats (the numbers and actual rogue cats are set by the teacher) are suspected of producing violent cats when crossed with any other type. New stories about Katzville, where this problem is particularly prevalent, were designed to be accessed via internet links reached from this screen. This data would be reviewed by the students in order to make predictions about rogue cats prior to the exercise. In the study these were paper based.

In this more untidy and less-clinical 'field research' setting, students observe tests of cats produced by crossing genetically-modified types. At the top of the screen is indicated the number of types suspected of containing rogue genes, such that, when crossed with any other, they produced a cat that **usually** becomes violent. For example, there may be five types of GM cat (A, B, C, D, E) producing cross-breeds of AD, CE etc. If only the original A cat contains the rogue gene, then AD would **probably** become violent but CE would **probably** not. The amount of uncertainty is set by the teacher. The uncertain link between cause and effect means that the outcomes of tests, even when the original cats with the rogue genes have been identified, are less predictable and these chief causes are more difficult to determine. The uncertainty is explained in terms of non-ideal testing conditions or nature-nurture issues.

As before, prior to each test, the left and right students enter their ideas individually as to their current theory about the rogue cat(s), and then their agreed idea. A cat produced by cross breeding the originals then appears on the test-rig for testing, apparently chosen randomly by the computer, and left and right students enter their individual predictions regarding the outcome. After then entering their agreed prediction, the development of the cat is accelerated in the RDA and the outcome (indicated by whether the cat becomes violent or not) is stored on

a whiteboard above the test bench. Since the development accelerator is reversible, no cat years of life are lost in this futuristic scenario!

The object of the mission is to run as many tests as needed (the students decide when to stop) to identify the original cat type(s) causing a problem. The number of tests carried out by students can be a matter of cost and the required amount of certainty. A prediction calculator is available from the mission testing screen that displays a running total of the amount of money spent (a product of the number of tests and a figure set by the teacher) and the percentage correct predictions made in the last X attempts – where X is student-defined. So, for example, students could decide to quit when they had 80% correct predictions for the last 10 predictions.

Feedback is only given about the distribution of joint opinions, thus if the players always select the answer of the person on the right as the shared answer, even if it differs from the player on the left, a message will appear asking the players to consider the others opinion. The system is measuring the amount of capitulation, capitulating three times more than your partner over the course of eight tests generates a message.

4.2.3 The filing cabinet

Although its use was not investigated in the study, the students can access a record of their inputs, the outcomes of the tests, and formative feedback from the training blocks either from the introductory page or when in the mission. The data recorded for the mission lists the cats crossed, the outcome, and whether their shared prediction was correct or not. If any feedback is given on working together this is also listed.

4.3 Description of additional resources and materials used in trials to support learning

Additional material was provided in the study to introduce Debating the Evidence, this enabled data to be gathered on current knowledge, and to act as contextual background for the Katzville mission. This paper-based rather than online information was used for the trials in order to facilitate evaluation. The paper-based versions were produced in the form of worksheets that students wrote on.

The introduction to Worksheet 1 stated that customers in four cafés had complained of food poisoning after eating two sweets. The council had gathered feedback from ten further customers from each café about the impact of the two sweets they had just eaten, ie whether they were they ill or not. The students were presented with this information and had to determine the sweet that caused the problems. In Café A it was one sweet with a direct covariance, ie eating that sweet always produced illness. In Café B it was two sweets with a direct covariance. In Café C it was one sweet with a degree of uncertainty between eating and becoming ill (in two pieces of evidence the cause-effect relationship was reversed). Finally in Café D it was two sweets with a degree of uncertainty (one piece of evidence reversed). The students were asked to say which sweet it was, why, and how certain they were of their answer using the following scale:

1. I am not sure at all
2. I think I am possibly correct
3. I think I am probably correct
4. I am fairly certain I am correct
5. I am absolutely certain that I am correct

This worksheet was designed to indicate the current level of understanding about interpreting covariation evidence and reasoning about the uncertainty arising from anomalous results.

The second worksheet, Worksheet 2, was designed to identify the students' understanding of identifying causes. It described the mission context, that is, the five genetically modified cats and the increase in violent behaviour. There were six articles from various people involved in the identification of rogue cats. Each gave reasons for their opinion and details about the work and their position. These are summarised in Table 2 below:

Name	Reasons that are trustworthy	Reasons that are untrustworthy
Bill Edwards – amateur scientist	- Has used evidence - Sample size (30)	- Bias due to unfortunate personal experience - Unqualified
Grandma Lil	- Has used evidence - length of observation	- sample received special care - bias due to personal interest - Sample size (5)
Peter Struddle, Perfect Puss Cats Ltd	- Has used some evidence - Sample size (40)	- infers 100% predictive certainty achievable - bias due to commercial interest
Bill Wills, Government official	- Has used some evidence - Provides result with uncertainty	
Petra Evans, Local councillor		- infers less than 100% predictive certainty invalidates research results
Anna Grimes, Anti Genetically Modified Cats Association	- Refers to other organisations	- Bias due to organisation belongs to - Unqualified

Table 3: Analysis of data in Worksheet 2

The students' awareness of how evidence was being used was scored by attributing one point for each appropriate insight and deducting one point for an incorrect statement, using Table 2 as a guide. The students were divided into pairs to use the software on the basis of matching this score within each pair. This data also provided information from which students could hypothesise about the rogue cats initially rather than guess.

4.4 Description of role of teacher and researcher in trials, and overarching approach to teaching and learning

Paul Howard Jones (PHJ) took the role of teacher-researcher for the two sessions of the trials. The class-teacher maintained a background presence, but did not generally intervene. Mary Ulicsak and Jing Lu observed and took field notes. Two further Futurelab staff provided technical support in terms of filming and recording.

The software was intended to be fully integrated into a scheme of work that involved sensitising students to issues regarding evidence and uncertainty. However, since the focus of these trials was upon evaluation of the software, efforts were made to eliminate discussion of such issues during the trials, except between collaborating partners when using the software. Although such control detracts from the ecological validity of the trials, it improves the possibility of identifying where collaborative use of the software, as opposed to teacher-learner interaction, has furthered understanding.

It was considered that the software would support learning by:

1. Encouraging collaborative discussion of evidence and uncertainty through individual and collaborative decisions in pairs.

2. In training: improving scientific thinking skills about covariation evidence through combined peer collaboration, prediction feedback and formative feedback on strategies and collaboration.
3. In mission: improving understanding of uncertainty through combined peer collaboration and prediction feedback in the mission.

5. EVALUATION OF DEBATING THE EVIDENCE

5.1 Participants

The students were from the top Year 8 science class at a voluntary aided comprehensive school. In this school 77% of the students received five GCSEs grade A-C in 2004, the national average was 53.7%. Fourteen boys and 13 girls attended both sessions, with another girl attending the second session only. The data used was from the 13 pairs that attended both sessions. The students were split into 14 pairs and given a unique identifier based on group, gender and number in pair, eg 3B1 refers to the first boy in pair 3, 3B2 refers to the second. The pairs were formed on the basis of the validity of the individual's reasoning and the number of justifications given for their belief in Worksheet 2. Students were ranked according to first the number of reasons that matched those of the researcher and then by the number of justifications given. Pairs with low numbers had fewer valid reasons and fewer justifications in general.

The five pairs that were filmed in the second session were chosen as they gave lots of reasons but few matched those identified by the researchers.

5.2 Method

The students attended two one hour sessions. These were led by Paul Howard-Jones (PHJ). The lesson plans for the sessions can be found in Table 4 and Table 5.

Goal		To identify students' current understanding of predictions and certainty
Equipment		Individual Worksheets 1 (Café Sweets) and 2 (Katzville News)
Timing	**Activity**	**Content**
5 min	Introduction	PHJ will explain who we are and overview of next two days
15 min	Worksheet 1 – Café Sweets	Individually and in silence identify which is the dangerous pudding in all four scenarios
2 min		Collect sheets and give out answers – *No discussion of why answers correct*
4 min	Mission context	PHJ explains mission (also on start of Worksheet 2)
15 min	Worksheet 2 – 6 reports on GM cats	PHJ asks them not to write anything yet, then reads each article in turn aloud, then ask them to **individually** identify: • Certainty (circle number) • Why sure • Why less sure Stress they can use that article or another as justification. Prompt for time left, 10 min, 5 min etc. At end ask them to decide which cat.

Table 4: Session 1 format

The data gathered is of the format:

- field notes
- video recording of entire class
- responses to Worksheet 1
- responses to Worksheet 2.

Goal		To identify whether software impacts their understanding of uncertainty in predictions
Equipment		Worksheets B – originals from previous lesson Computers with DtE software running
Timing	**Activity**	**Content**
5 min	Form pair groups	PHJ asks the students to find their allocated partner and sit at that computer number
5 min	Recap	PHJ recaps task, ie training and cats
5 min	Overview of training	PHJ explains training using Data Projector
5 min	Overview of mission	PHJ explains mission including prediction calculator, costs, and filing cabinet
10 min	Training	As pairs identify genetically modified cat
10 min	Mission	Identify mutant cats – talking only in pairs
1 min	Review worksheet	Confirm answers B and D. Give out original Worksheet 2, ask them to look back and ask them to review what they wrote yesterday, has their trust changed? Did they miss anything they should have spotted?
10 min	Review hypotheses on Worksheet 2	Put a star by any changes and a description of why they changed their mind (either in margin or number and continue on back)

Table 5: Format of Session 2

The data gathered is of the format:

- field notes
- video recordings of five pairs
- revisions to Worksheet 2
- log files.

At the end of the session three students were interviewed by PHJ for approximately five minutes to identify their understanding of covariance and the task.

5.3 Approach to analysis

The students did not have dual key control when using the software, they both used the same inbuilt mouse on the keypad for entering their responses. Also, the formative feedback was limited to textual messages rather than allowing for pictorial prompts as originally planned. This did not invalidate the research questions, but limited the amount of comparison that could be made. Table 6 summarises what data was used to address each question identified in Section 0.

Research question	**Evidence used**
How does the proposed framework support the development of collaboration?	• Changes in individual and agreed computer responses taken from log files • Time taken for decisions to be made taken from log files • Video and field notes

Does the type of collaboration supported by the proposed framework transform the student's understanding of scientific thinking and uncertainty?	• Responses and revised responses to Worksheet 2 • Changes in individual and agreed computer responses from log files • Responses to interview questions
What is the role of feedback in different modes on motivating and engaging students?	• Video and field notes • Training and mission log files • Responses to interview questions

Table 6: Chief sources of evidence used to assess 'Debating the Evidence' software

5. 3 Results from trials

5.3.1 Session 1: Identification of pairings for second session from Worksheet 2

Table 7 shows the score, which is the number of reasons given agreeing with those identified by the researchers, and the total number of reasons given after the first session. For example, the statement: "if it was wrong it would have been very mean on all cats and she has no real evidence and no test or research results to back her up" or "the goverment dismissed and I don't believe the goverment" would be two reasons, the statement: "none" would be zero.

Identifier	Score	Total number of reasons
1B1	-1	4
1B2	0	7
2B1	0	8
2B2	0	7
3B1	1	14
3B2	0	4
4B1	1	2
4B2	2	13
5G1	3	10
5G2	2	13
6G1	3	12
6B1	3	13
7G1	4	12
7G2	4	16

Identifier	Score	Total number of reasons
8G1	4	9
8B1	4	11
9G1	5	11
9G2	5	18
10B1	6	17
10B2	5	6
11G1	7	11
11B1	7	15
12G1	8	13
12G2	8	15
13G1	9	15
13B1	9	12

Table 7: Pairings of study participants

These worksheets were amended in the second session.

5.3.2 Session 1 and 2: Initial understanding of covariance and amendments to Worksheet 2

For each student, the change in confidence when the cause-effect relationships became more inconsistent was calculated by adding confidence for problems A and B in Worksheet 1 and subtracting from this the combined scores for C and D. A negative score here indicates an appropriate awareness of how confidence may be reduced by anomalous results. This is shown in Table 8, together with the number of additional insights that students were able to make about using evidence after experiencing the software. Finally, based on the number of

additional insights made the students were divided into three groups depending whether the insights were gained by neither, one or both group members, these groups were classified as follows:

A. Pairs in which both gained some further insights (five pairs)
B. Pairs in which one individual gained further insights (five pairs)
C. Pairs in which neither individual gained further insights (three pairs)

10B1 and 10B2 could not be included in the study as their log file was corrupted.

Identifier	From Worksheet 1: Confidence scores (A +B) – (C+D) (change in confidence with uncertainty)	From Articles in Worksheet 2: Extra insights gained after using software	Group allocated to based on extra insights gained
1B1	-3	0	C
1B1	3	0	C
2B1	-1	0	C
2B2	1	0	C
3B1	-2	4	A
3B2	0	4	A
4B2	-2	1	B
4B1	0	0	B
5G1	-3	0	B
5G2	-2	1	B
6G1	-3	1	A
6B1	0	4	A
7G1	-3	2	A
7G2	-3	6	A
8G1	-6	3	A
8B1	-1	1	A
9G1	0	0	B
9G2	1	4	B
10B1	-2	2	A
10B2	0	1	A
11G1	-2	0	C
11B1	0	0	C
12G1	-1	6	B
12G2	0	0	B
13G1	-2	2	B
13G2	-2	0	B

Table 8: The second column of this table shows the change in confidence indicated by students when dealing with cause-effect relationships that were inconsistent rather than consistent. A negative score here indicates an appropriate awareness of how confidence is related to consistency of evidence

When using the training software, ten of the 13 groups used a mature strategy as individuals and pairs. Those that changed strategies are shown in Table 9.

Identifier	Attempt 1 Strategy	Joint strategy	Attempt 2 Strategy	Joint strategy
4B1	Pattern matching	Pattern matching	Mature	Mature
4B2	Pattern matching	Pattern matching	Mature	Mature
6B1	Pattern matching	Pattern matching	Mature	Mature

6G1	Pattern matching	Pattern matching	Mature	Mature
8B1	Vacillating	Mature	This pair was falling behind, so PHJ sent them on to the mission.	
8G1	Vacillating	Mature		

Table 9: Training feedback received when not initially mature

6. ANALYSIS

This section addresses the initial three research questions in order.

6.1 How does the proposed framework support the development of collaboration?

Once the causes had been identified in the training, the students did not appear to collaborate further. Instead, they just chose the correct solution, often with one student doing all the selecting. This was observed and recorded by the researchers in their field notes, filmed on video and can also be deduced from the rapid reaction times that occurred after the first few successful predictions with correct theories.

This contrasts with the mission where for Group A, in which both gained some further insights, delays (associated with reflection) occurred **after** some initial rapid responses for the four pairs where data was gathered. This appears to demonstrate a continuing reappraisal in the face of occasional unexpected results. Unfortunately, the video film was not of sufficient quality to allow transcription of the dialogue. However, it does support observations recorded in the field notes that, unlike towards the end of the training, students were entering their individual answers and generally taking it in turns to enter agreed responses. Also, it showed the students were making considerable use of the prediction calculator, even though the software required this to be opened and closed on each occasion of use.

The individual responses provided during the mission are helpful in determining whether collaboration, in terms of the co-operative sharing and testing of ideas, was developing during the use of the software. These responses were recorded in the log files that were used to generate Figures 3 to 7. These files recorded individual and agreed theories and predictions, and the times taken to produce them. One example that may demonstrate a developing co-operation in the testing of ideas is shown in Figure 3 where discrete theories are initially held by 3B1 and 3B2, before an agreed 'best fit' is settled upon in test 5. When this pair next departs from this theory (which was only partially correct), they abandon mutual theories and experiment with new ideas together. (NB In the training causes differed between student pairs and were randomly selected by the computer. In the mission, however, the causes were B and D for all students although data sets were still unique due to the randomly distributed anomalies.)

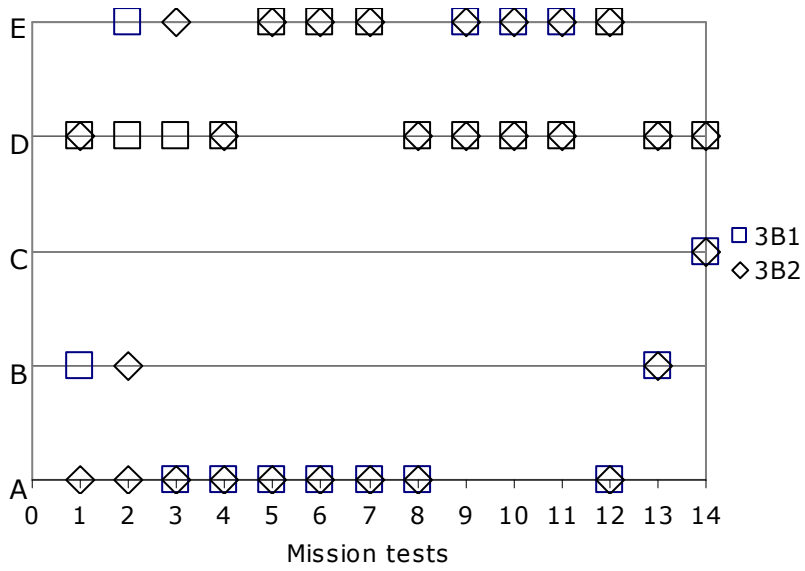


Figure 3: The theories of 3B1 and 3B2. This pair originally determined their ideas individually before converging on a common theory. When they next depart from their theory (which was only partially correct), they simultaneously abandon their mutual theories and experiment with new ideas together.

A similar pattern is seen with 6G1 and 6B1. They appear to hold different theories until the fourth test, and then become consistent and choose the pair of genetically modified cats that are more likely to be violent:

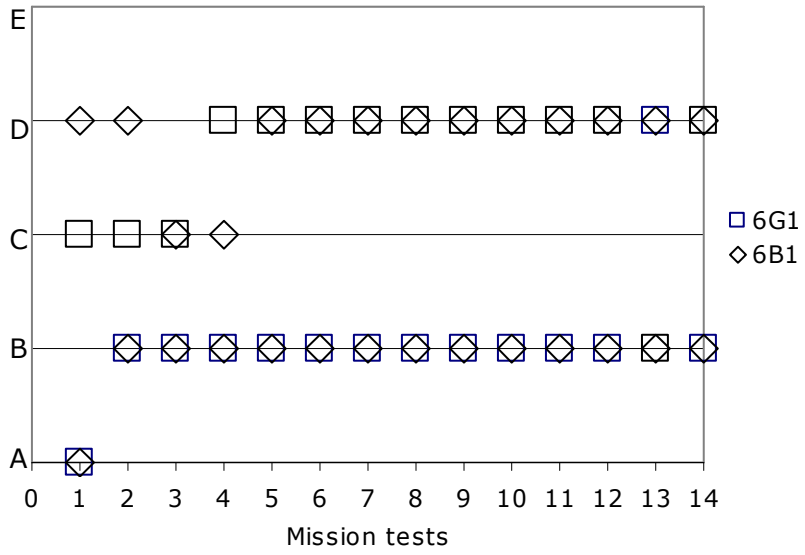


Figure 4: The theories of 6G1 and 6B1 showing a convergence of partner's opinions from test 5

In Figure 5, 8B1 and 8G1 disagree originally then converge onto the same two solutions (one of which is not correct and not giving optimum prediction success). However, one decides to stick while the other departs to an alternative pair of causes (neither of which is correct). This student soon returns to the partially successful hypothesis. The noteworthy thing here, however, is that this idea was only entered once as the agreed response, suggesting that co-operation was occurring in terms of testing individual hypotheses.

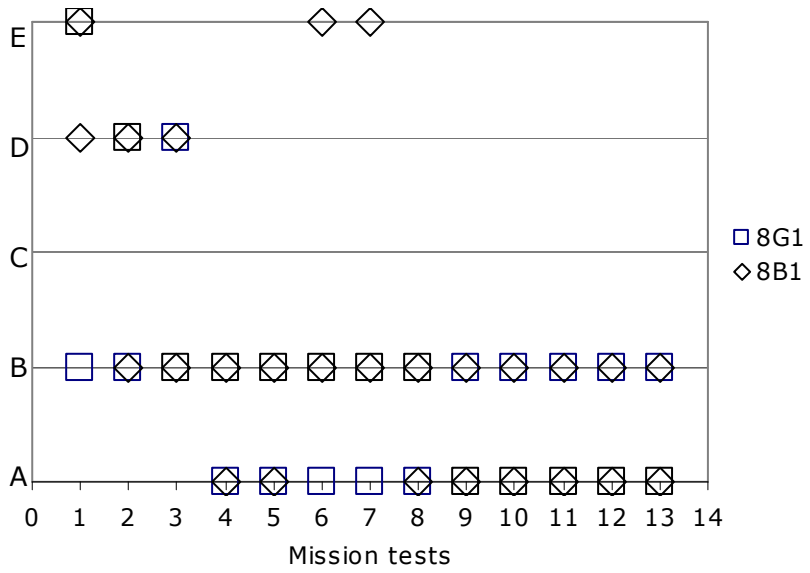


Figure 5: After the theories of 8B1 and 8G1 converged, one partner considered an alternative for a couple of trials before deciding to stick, with their partner, to the same two solutions.

7G1 and 7G2, the last pair in Group A for which there are log files, showed a similar pattern of co-operation.

Other pairs, even those not showing any later development of insights (eg Group C) also demonstrated a sharing in terms of contributing their individual ideas as agreed predictions, as in 1B1 and 1B2 who eventually reached the most appropriate two solutions as shown in tests 22 and 23 in Figure 6:

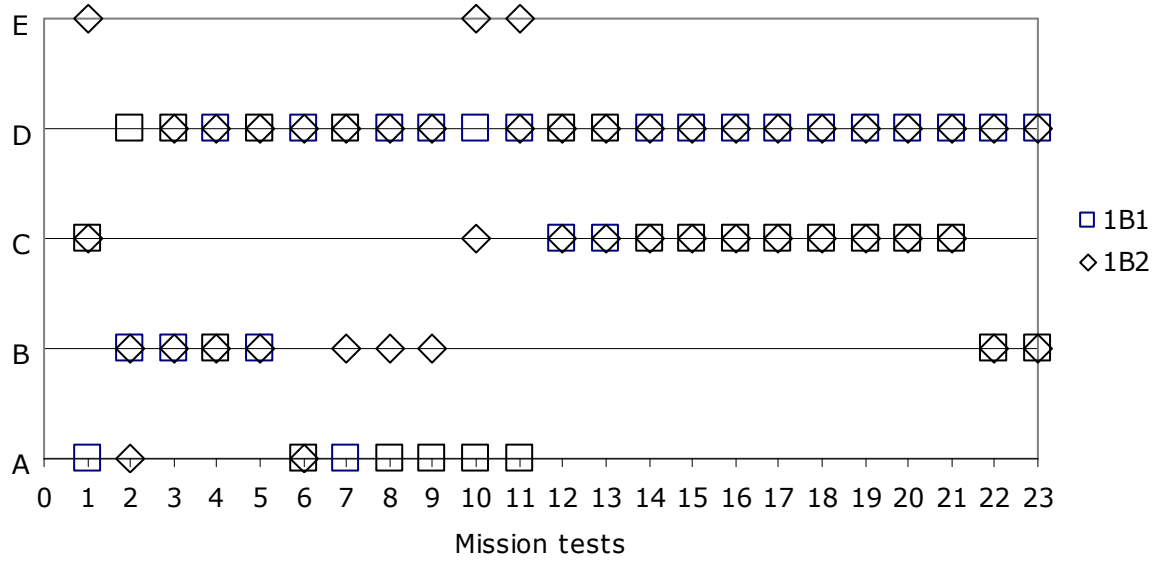


Figure 6: 1B1 and 1B2 eventually reached the most appropriate two solutions

From analyses of this type it appears that partners were cooperating with each other and that their decisions were leading towards agreed optimal solutions. This suggests that collaboration is being supported by the system but it should be noted that further qualitative research required to corroborate this finding.

6.2 Does the type of collaboration supported by the proposed framework transform the students' understanding of scientific thinking and uncertainty?

From Worksheet 1, the students' scientific thinking about covariance evidence was good. Only one failed to spot the causes in café problems A and B dealing with completely covariant cause-effect relationships. The café C and D problems involved uncertainty in the cause-effect relationship, but all students spotted the single cause in C and at least one of the two causes in café problem D – although four students did not identify both causes correctly. However, despite this success in identifying causes, 10 out of the 27 pupils did not decrease their level of confidence for answers when anomalous data were present. That is, their confidence in C and D relative to A and B was not lower, indicating less understanding amongst these pupils for the relationship between scientific thinking and uncertainty.

When analysing the additions that students made to Worksheet 2 (the articles), after using the software (these results are shown in Table 8), researchers did NOT include students' own conclusions about which were the causes - even if these were evidence-based. The only insights that were included in the score were those made about the validity (or otherwise) of the evidence basis of the articles. For example, 9G1 was less sure about Grandma Lil's conviction that B cats were fine, because her own evidence was pointing in a different direction: "My results show that the B cats have the genes that make them rough…". This type of comment was **not** included as an additional insight about the use of evidence, whereas spotting that Grandma Lil's sample size was very small and probably not typical was counted. The comments that were counted as additional insights only included factors already present in the article that gave grounds for confidence or otherwise in the evidence, although these may have arisen from the student's experience of collecting their own evidence. For example, after using the software, 9G2 commented that she was less sure about Peter Struddle (Perfect Puss Ltd)'s claim that less than 100% predictive success meant that the evidence could be dismissed: "Getting 100% predictions is very difficult. It would be useful to see his results, and see which cats reacted".

Overall there was an increase (about a third again) in the number of valid comments made about the way in which characters featured in the articles had used evidence to support their arguments. More gains were made by students who had already demonstrated in the previous lesson some grasp of how inconsistent evidence weakens confidence in findings. Indeed six out of the 10 students (60%) who had failed to demonstrate this understanding in the preliminary lesson did not make additional insights. Only four out of the 17 who had previously demonstrated this understanding did not make further insights (23.5%).

Although no other pedagogic strategy was used, a direct link between this apparent increase in insights and using the software is difficult to prove in absolute terms. Interviews with students who had been filmed were unsuccessful in exploring their reasoning about uncertainty and the reasons behind their decisions, despite evidence from their computer responses that they had successfully reasoned out the best-fit solutions from data that contained anomalies. This may be because, especially in school and in front of peers and visitors, confidence is strongly associated with achievement – ie it is suggested students feel they *should not* be uncertain and feelings of not-knowing tend to prompt silence. However, in trying to link the increase in insights with using the software, it is possible to make a relationship between their experience with the software and outcomes in terms of increased insights. In attempting this, the experience of each pair must be considered individually since, due to the random elements built into the software, the tests and outcomes observed, as well as the progress made, were very different for each pair.

The experience of Group A, in which both individuals made progress with their insights, had experiences in which some advantage appeared associated with an appropriate latency in their response to incorrect predictions. That is, they were rewarded for changing their theories in response to a series of incorrect predictions. 6G1 and 6B1 (shown in Figure 4) changed their theory initially in the face of evidence but, after some success, then held onto the correct theories despite two occasions of conflicting evidence. 8B1 and 8G1 started off with the two

correct theories but then gave up one in the face of some mixed prediction results, but they then hung on to another incorrect theory, A (see Figure 5). However the log files showed this was reasonable – as they were still enjoying a period of good prediction success. 7G1 and 7G2 (who carried out 24 tests) also changed their ideas initially and then settled on the correct solution, experiencing general success in their predictions. However, in addition to one conflicting prediction, as if testing whether prediction success was a solid indicator of having identified cause, they also momentarily changed their theories twice and still got correct predictions. 3B2 and 3B1 (see Figure 3) never arrived at the right two causes, but showed an appropriate latency in modifying their ideas, as if having grasped the notion that, in this more real world, an incorrect prediction no longer meant you had to abandon your theory immediately.

This is in contrast to the experience of Group C, in which neither individual made progress with their insights. These pairs either ignored disconfirmatory evidence for too long or, by chance, did not encounter enough anomalous outcomes to challenge their ideas. 1B1 and 1B2 (23 tests shown in Figure 6) revealed a good example of how prior belief can hamper progress, hanging on to an incorrect theory despite prediction success at less than 50%. 11G1 and 11B1 (nine tests) identified the correct two causes and then enjoyed perfect prediction success. 2B1 and 2B2 (11 tests) found the two causes early on and experienced only one conflicting prediction outcome.

Thus we can conclude that the software does impact **some** students' understanding of scientific thinking and uncertainty. This is without scaffolding from the teacher or discussing with students other than one's partner. However, it does not always lead to an improvement, as shown by those who do not change opinions regardless of evidence presented. As with the first research question, without a dialogue-based study, we cannot accurately characterise the quality of the collaboration that the software supports since, for example, a pause between decisions shown in the log files does not imply discussion about opinions. This implies we cannot make definite statements about the impact of collaboration on understanding in this context.

6.3 What effect does the feedback, presented in different modes, have on motivating and engaging the students?

The system provided two types of feedback:

Prediction feedback: The students witnessed the outcome of each test and were thus made aware of whether they had correctly predicted or not. In the mission, the prediction calculator allowed a prediction success (as the percentage correct in the last N attempts, where N could be set by the student). In training, the percentage correct for the block was provided.

Formative feedback on collaboration and thinking strategies: In the training lab, formative feedback and advice on problem-solving and collaboration strategies was provided, as appropriate, at the end of each block. In this study, only one training block was set, so such advice would only appear if the students failed their training. As discussed in Section 0 originally it was hoped that this could be provided in diagrammatic and text form, but budgetary constraints prevented this feature being included in the prototype software.

6.3.1 Effect of prediction feedback

In their first attempt at training, most of the pairs (apart from groups 4 and 6) were motivated by the results of the tests to revise their theories until they achieved sustained prediction success, thus successfully identifying the two causes (a so-called 'mature' strategy). Figure 7 illustrates a typical mature problem-solving strategy as the agreed responses during training (for 8G1 and 8B1) move towards the solutions (shown by dotted lines).

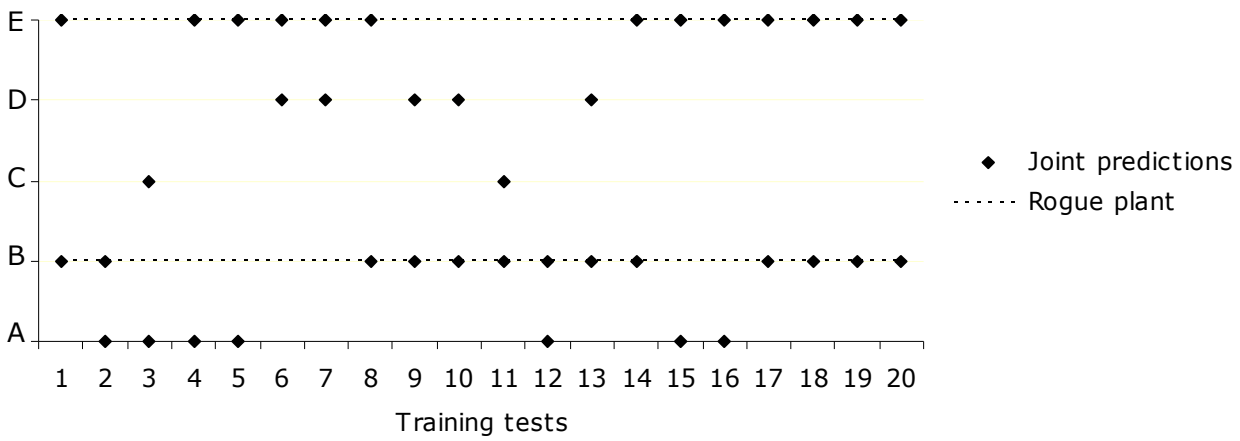


Figure 7: A typical mature problem-solving strategy during training (8G1 and 8B1), as characterised by the agreed responses converging upon appropriate solutions (shown by dotted lines) that give 100% prediction success.

So, prediction feedback clearly orientated students towards identifying causes, and provided motivation up to, but not far beyond, the point at which these causes were identified.

The students who used a mature strategy from the beginning did not appear motivated by the training, regardless of feedback, once the two causes had been identified. This was observed and recorded by the researchers in their field notes, filmed on video and can also be deduced from the rapid reaction times that occurred after the first few successful predictions with correct theories.

6.3.2 Effect of formative feedback

Two of the pairs (6B1, 6G1 and 4B1, 4B2) failed their training and were diagnosed by the computer as following a strategy of pattern matching. In line with this diagnosis, their TPC was poor but they were still scoring above chance predictions when the combinations were repeated in the second half of the block. Appropriate feedback and advice was provided by the system and these students' thinking strategies became mature in the second attempt.

In the absence of any teacher/researcher intervention, from these two pairs it appears that the feedback from the computer supported the development of the students' problem-solving strategies. However, it has been shown elsewhere, that the strategies of students working individually with these types of problems can be progressed by prediction feedback alone. There is also some evidence from previous research (Howard-Jones et al 2005) that this progression may be further influenced by collaboration. Further research would, therefore, be needed to identify the individual contributions to the students' learning made by the prediction feedback, the automatic formative feedback and collaboration, and to explore the interaction between these. Similarly, it would be useful to explore whether changing the format of the feedback, for example, to a graphical representation of the strategy recognised, or even a shortened text message, would benefit the student's understanding and motivation.

In this high ability class, only two pairs could have benefited from the formative feedback, although their improvement does provide some tentative evidence that the formative feedback is helpful.

It should be emphasised that the evaluation of the Debating the Evidence software was quasi-experimental and the software was not evaluated as part of an integrated work scheme. That is, all additional instruction and scaffolding of the students' thinking was deliberately and carefully avoided by the teacher-researcher. (For example, although the software provides

excellent opportunities for classroom discussion and leading the pupils thinking through questioning, none of these were exploited.) This served to focus the evaluation upon the software, but probably minimised the overall learning benefits that the software offers when fully integrated in pedagogic terms.

7. GENERAL FINDINGS

In this section the general findings are discussed. These arose from the various stages of the software development, the field notes and log files, and are not necessarily findings relating to the research questions. This section includes possible directions for future work.

7.1 Educational research community

The evaluation provides some evidence that merely interacting with a computer simulation of a scenario where this is a less than perfect cause-effect relationship does improve students' ability to sceptically examine how evidence is used. Further evaluation with larger sample numbers and control groups is needed to confirm this.

There is some evidence that rapid feedback in the form of prediction outcomes engages and supports students in revising their theories until the correct solutions are reached. This is shown by the development of mature strategies in the training blocks.

Asking for individual as well as agreed responses allowed the system to monitor inter- and intra-individual performance, as well as allowing individual explicit expression of views. The pupils appeared motivated to do this, and may even have been motivated **by** this in their collaboration, but **not** when the decisions became routine. Then it seemed to automatically befall just one person to enter all responses.

The experience of using the software caused some students to progressively adapt their thinking strategies, without intervention from a teacher. Also, after using the software, most groups were able to identify additional issues when allowed to augment their previous attempt to critically appraise the use of evidence in some newspaper reports. The joint effect of using peer collaboration, prediction feedback and formative feedback in a simulated encounter with scientific uncertainty undoubtedly supported this gain. However, a larger study might valuably identify the individual contributions made by these factors to the learning gains that were apparent.

Although completing the training facilitated the dynamic analysis of the students' strategies and thus the provision of formative feedback, the students' motivation dropped during training as soon as they were confident that they had identified the two causes. This contrasted with the mission, where the element of unpredictability, even when causes had been correctly identified, maintained the students' interest in sharing opportunities to make further predictions and increase the evidence base.

The software appeared successful in engaging students' interest in uncertainty and in supporting collaboration that produced additional insights. Further research is needed to confirm these suggested benefits. Some of this research needs to include the evaluation of educational interventions that integrate the software with standard teaching techniques. The following further questions have also been raised by this study:

1. How do students discuss uncertainty and does this vary with age? What are the barriers?
2. What are the different ways that students interact when solving problems containing an intrinsic element of uncertainty? How are these characterised and how do they relate to performance?

3. Does uncertainty itself motivate? If so, why and how can this be exploited?

7.2 Teachers, advisors and head teachers

The interactive encounter with simulated scientific uncertainty engaged students with the realities of evidence in a real world and suggests this can improve their critical consideration of how evidence is used. This could be a vital way of improving students' critical awareness of the role of evidence in 'science in society' issues. The heterogeneity of outcomes within the class and the diverse experiences that different pupils had with the same software also suggests that the experience could be a useful precursor to classroom debates about the importance and limitations of evidence.

Students seem to find uncertainty both perplexing and engaging. Worksheet 2 indicated that many students believed the amateur scientist was incorrect because he could not add up, this was incorrect as the discrepancy may have been due to breeding two rogue cats. This suggests that awareness needs to be raised, despite uncertainty being discussed in the introductory lesson. This misunderstanding was also indicated by the surprise and interest generated by the emergence of a self-conflicting data set. Students did not find it easy to discuss uncertainty, perhaps because uncertainty in the classroom is usually associated with the risk of failure. Few pupils feel good about not knowing an answer. Engaging students with discussions about uncertainty may be vital in improving future science-in-society debates, and this type of simulation software may contribute here. It would also support the 21st Century Science curriculum, as it focuses on understanding and analysing of data rather than the acquisition of facts

7.3 Policy makers

Many of the students were ready to dismiss one opinion featured in the article simply because the speaker was old and appeared frail, for example, 9G2 reasoned: "Elderly people sometimes forget" while 1B2 believed: "I think she has posibly lost her marbels". The students were also more ready to dismiss the views of businesses and government on the basis of suspecting their prejudice rather than on the basis of their evidence, ie in preference to factors such as sample size, uncontrolled influences, overly certain findings, qualifications, experience and other issues. This unreasoned mistrust of authority was often prominent in the responses of the students, echoing some of the popular rejections of government advice regarding recent issues such as MMR. This may reflect a need for greater attention to be given in the curriculum to the critical analysis and appreciation of scientific evidence in areas of social concern.

7.4 Software developers

Students appeared motivated by interactive software that allowed them to make predictions about problems containing imperfect cause-effect relationships (although this may depend on responses to the educational research questions listed in Section 0). This was shown by the contrast in attitudes between completing the training and collecting evidence on the mission. This supports the notion that risk-taking in virtual scenarios is one of the major factors explaining the high degrees of motivation and engagement provided by some computer games (Gee 2003). The deliberate introduction of risk and uncertainty is generally alien to education, but may be a way of improving students' engagement with educational software.

Further development of the prototype software for commercial production might also take note of the following points:

1. Film evidence shows that the prediction calculator was often used, presumably as a means to monitor progress. This supports the opinion, expressed by the two class

teachers at the usability trials, that any future version of the software should keep this continuously visible during the mission to provide a more 'game-like' quality to using the software.

2. The training becomes tedious when the students are performing well, and a means to curtail this as soon as a mature strategy is reached would be of benefit.

Further work is required to investigate whether dual mouse control of a single cursor is an effective method of providing dual control without one player monopolising the inputs.

REFERENCES

Gee, JP (2003). *What Video Games Have to Teach Us About Learning and Literacy*, Palgrave Macmillan

Howard-Jones, P (2004). *Debating the Evidence: Social Evidence-based Decision-making* (SED), submission to CFI

Howard-Jones, PA, Joiner, R, Bomford, J (in press). *Thinking with a theory: theory-prediction consistency and young children's identification of causality*, Instructional Science

Ross, L (1977). *The intuitive psychologist and his shortcomings: distortions in the attribution process.* In Berkowitz, L (ed) Advances in Experimental Social Psychology, 10, New York: Academic Press