Changes in Student Decisions with *Convince Me*: Using Evidence and Making Tradeoffs

Marcelle A. Siegel (mcgull@socrates.berkeley.edu)
Science and Math Education (SESAME); 4533 Tolman Hall
University of California at Berkeley
Berkeley, California 94720-1670 USA

Abstract

This study examined the cognitive processes of decision making in an urban high school classroom in which tenth graders analyzed scientific evidence about current issues of technology and society. A computer program, called *Convince Me* (Schank, Ranney & Hoadley, 1996), provided scaffolding for making evidence-based decisions for the experimental group. During the course of instruction, both the control and experimental classes completed open-ended assessments. Student progress, in using evidence to support claims and in weighing benefits and drawbacks, was mixed. Reasons for the changes in decision making are offered.

Coherent Reasoning about Evidence

This research emphasizes reasoning skills in using evidence. Other studies have examined these skills as well. For example, in the Knowledge Integration Environment (KIE) project, students interpret and critique scientific information garnered via the internet and make conjectures about it, forming a scientific argument. KIE researchers have hypothesized that engaging students in the creation of an argument facilitates conceptual change (e.g., Bell & Linn, 1997).

The program used in this project,*Convince Me* (CM), possesses a connectionist network (called ECHO) that simulates human reasoning. Using ECHO, CM offers feedback as to whether the student's evaluation of each proposition matches the values that are simulated by the computer. ECHO's principles of reasoning are based on the Theory of Explanatory Coherence, established by the philosopher, Thagard (1989). The theory assumes that the plausibility of a belief increases with, for instance: a) the simplicity with which it is explained, b) increasing breadth of evidential coverage, and c) decreasing competition with alternative beliefs (Ranney & Schank, 1998). Table 1 lists these principles in more detail.

While using the program, a student enters alternatives, beliefs, and evidence about an issue and then evaluates the plausibility of his decision. CM's interface provides scaffolding for making a decision through prompts for entering hypotheses and evidence. Students are also asked to make links between and among hypotheses and evidence and must choose whether each link they make is supportive ("explain") or contradictory ("conflict"). Students rate the reliability of each piece of evidence, as well as how much they believe each statement that they have entered. Next, they run the ECHO simulation, and then they contrast their ratings with ECHO's activations by pressing the Model's Fit button that calculates a correlation score and responds with, for instance "The correlation between your ratings and ECHO's evaluations is: 0.29 (mildly related)...."

Table 1. Some of ECHO's Principles for a Coherent Argument

1)	Plausibility increases with more support from explanatory statements.
2)	Plausibility increases with less competition from contradictory statements.
3)	Simplicity: The plausibility of a belief is inversely related to the number of cohypotheses it needs to explain a proposition.
4)	Data priority: Results of observations, such as evidence and acknowledged facts, have a degree of acceptability on their own.

Prior studies with CM indicate that it is a useful tool for learning about reasoning. Students using the program performed better than students doing similar pen and paper exercises, perhaps because of the computer's feedback (Schank, 1995). Also, undergraduates working with CM improved at distinguishing between hypotheses and evidence (Ranney, Schank, Hoadley & Neff, 1994). High school students using CM supported their beliefs with objective evidence and generated more than one alternative while making complex decisions (Siegel, 1997).

Research Focus

This study investigated whether using *Convince Me* with high school students in an issue-oriented biology class helped the students become better at using scientific evidence to support their decisions and to weigh tradeoffs in their choices. The hypothesis was that CM's principles of coherence (Table 1) would help engender better decision-making skills.

To test whether CM activities significantly improved students' uses of evidence and the weighing of tradeoffs, two classes were examined for several months. The researcher observed, taught, participated with, and tested the classes daily and equally from January through June of 1998. The same teacher led both advanced Biology classes. The main curriculum was *Science and Sustainability*, a new course developed by the Science Education for Public Understanding

Program (SEPUP) at the University of California at Berkeley's Lawrence Hall of Science. SEPUP's courses include written materials for students and teachers, as well as laboratory equipment. Students learned science by studying, discussing, debating, and experimenting, based on issues relevant to society.

Assessing Student Reasoning

In addition to SEPUP instructional materials, this study utilized an assessment system developed by SEPUP. The system was developed for SEPUP's middle school course using Rasch measurement techniques (e.g., Masters, 1982).

The SEPUP assessment system consists of *variables* that are a set of scientific concepts, processes and skills that are central to the course (Roberts, Wilson & Draney, 1997; Sloane, Wilson & Samson, 1996). *Understanding Concepts* and *Evidence and Tradeoffs* were the variables used in this study. Items for measuring these variables were embedded in tasks throughout the curriculum. Students wrote short essays or sentences in response to open-ended questions. Both Evidence and Tradeoffs and Understanding Concepts questions were scored on a criterion-referenced, 0-4 scale according to the SEPUP rubric (0 is low, 4 is high: see Roberts, et al., 1997).

Students were assessed on two elements of the variable Evidence and Tradeoffs:

- *Using Evidence* (student supports claims with relevant evidence)
- *Using Evidence to Make Tradeoffs* (student sees drawbacks as well as benefits in choice and supports these tradeoffs with evidence)

For example, a student who provided the major objective reasons for her decision and supported each reason with relevant and accurate evidence would receive a score of 3, on the 0-4 scale for Using Evidence.

Students were also assessed on two elements of the Understanding Concepts variable to determine how well they grasped the principles of coherent decisions:

- *Recognizing Relevant Content* (identify and describe the principles of coherence used in *Convince Me)*
- *Applying Relevant Content* (use the principles of coherence in new situations)

In accordance with the SEPUP method (Sloane, et al., 1996), students were introduced to the assessment system before the evaluation took place. They completed practice questions and received feedback. They also used the scoring guides while constructing their responses in order to learn to distinguish the qualitative differences between score levels.

Participants

The school involved in the project faced socioeconomic challenges typical of the inner city. According to the school district's data, 50% of students qualified for Aid for Families with Dependent Children (AFDC); 42% of students were identified as Limited English Proficient. The most recent SAT scores reported for the approximately 41% of school seniors who took the test were far below the national average: 321 on the verbal portion (national average is 423) and 437 on mathematics (national average is 479).

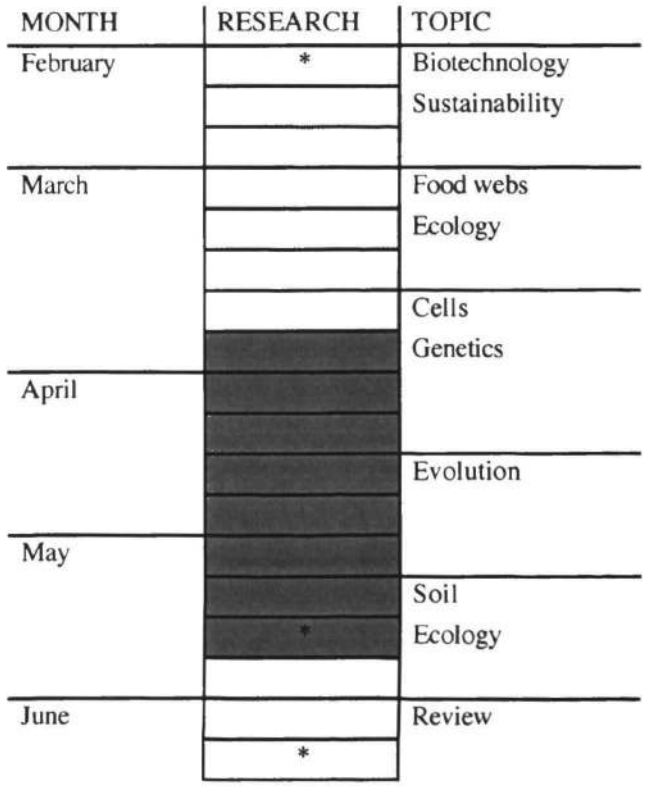
Before the study began, the two classes were compared to see if there were initial differences in ability. Scores on the standardized Terra Nova test (CTB, 1997) were obtained. The SEPUP control class had an average reading performance level on a 1-5 scale (5="advanced" and 3="nearing proficient") of 2.34 (average percentile of 60.25), while the experimental class had an average reading performance level of 2.56 (average percentile of 67.25). However, a t-test revealed that the difference was not significant (p=.35).

Procedure

Both of the two tenth-grade classes participated in the *Science and Sustainability* course activities. For a period of two months, the experimental class used both the course and the CM computer activities. Due to scarcity of computer facilities, half of the class would use CM one day while the other half completed SEPUP activities; the next day they would switch. The control class engaged in SEPUP decision-making activities during the time the experimental class used CM in this manner. The timing of the activities, the evaluations and the topics covered are summarized in Figure 1.

Figure 1: Timeline

The 2nd column is separated into rows with each representing 1 week. The light areas represent weeks when both classes were working on the same activities. The darkly shaded area represents the experimental period. The asterisks indicate times of testing (except that the Evidence and Tradeoffs tests for Rasch analysis lasted three weeks).



Some of the questions used on the Evidence and Tradeoffs evaluations are shown in Table 2. Ten open-ended items were taken by 56 students in the two classes during three weeks at the beginning of the study, and ten open-ended items were taken over three weeks after the decision-making activities at the end of the study. In addition, five Understanding Concepts questions (examples in Table 2) were given to the experimental class immediately after the *Convince Me* activities.

Table 2. Examples of Evidence and Tradeoffs and Understanding Concepts Questions

Evidence and Tradeoffs Items	**Question #**
Do you think humans should be included in the Antarctic ecosystem? Why or why not? What role, if any, do you think humans play in the Antarctic ecosystem?	3, 4
Imagine that you are the principal of a school that is having a problem with broken windows. Your choice is to replace the broken windows with either glass or plexiglass (plastic). Glass and plexiglass have different properties, and the plexiglass costs about 25% more. What material would you use and why? Be sure to describe the trade-offs involved in your decision. A complete answer will discuss the advantages and disadvantages of both materials.	7, 8 also 19,20
...Table of Final Radish Heights (Calculate the average height for each treatment.) a) Would you add fertilizer to soil to increase agricultural output? Give reasons for your answer. b) Do you think that adding fertilizer to soil to increase agricultural output is a sustainable process? Explain.	13, 14
Understanding Concepts Questions	
H1: Diazinon should not be used because it is dangerous to animals E1: It causes birth defects in chickens (Reliability=3) E2: It poisons and kills birds, bees, and fish (Reliability=3) vs. H2: Diazinon may be used because it is safe E3: It does not cause birth defects in rats or rabbits QUESTIONS ABOUT THE ABOVE ARGUMENT: 1. Which side of the argument, H1 or H2, would ECHO think is stronger? Why? 2. How would you change the argument to make H1 stronger?	4,5

Analyses

Two types of analyses of the data were carried out. The first analysis (*basic*) included data in raw form, and looked at particular points in time (*tests*). In this basic analysis, the first question (***pretest***, week 1) was compared to the question given first after the treatment period (***posttest***, week 16) to the last question given (***delayed posttest***, week 18). While this type of testing might be more familiar to researchers and classroom teachers, this type of analysis is not fully sound according to theories of measurement.

For example, the researcher has no way to tell which items are more difficult than other items. To address such measurement issues, a second analysis was carried out using Rasch modeling (e.g., Masters, 1982; Wright & Masters, 1982). Using this advanced statistical technique, one can compare the difficulty of items in detail, rather than giving questions without knowing how they differ, because "item difficulty" and "person ability" are estimated on the same scale. In addition, Rasch modeling is not only norm-referenced (i.e. comparing a student to other students), but also criterion-referenced (i.e. comparing a student's work to content standards or criteria) which offers more avenues for interpreting the results.

In the Evidence and Tradeoffs Rasch analysis, two sets of data were modeled: all the items answered before the treatment (***preexam***) and all the items answered after the treatment (***postexam***). Because the questions were open-ended and were often embedded in a laboratory activity, multiple questions could not be given on the same day; the preexam and postexam both took three weeks to complete. Students' progress over time and difficulty of items were thus confounded. In order to control for the difficulty of the items, a simulation was run with items anchored to an analysis of 830 students in another SEPUP course (which was possible because there were common items) (Wilson, Sloane & Roberts, 1995). Four preexam items and two postexam items were anchored at their appropriate difficulty levels from this previous study. In the Understanding Concepts analysis, the five items were modeled as one (post) exam. All the student scores were analyzed using *Quest* software (Adams & Khoo, 1993).

The ET and UC questions had been pilot tested for reliability and validity in a previous study. Using this information, ET items were assigned to the two exams.

Rasch Results: UC Correlation

Results from the Understanding Concepts evaluation suggested that understanding the principles of *Convince Me* was helpful in building better decisions in Evidence and Tradeoffs. Students who did better on the Evidence and Tradeoffs postexam also did better on the Understanding Concepts exam. The correlation between ET and UC was .50 for the basic analysis and .61 for the Rasch analysis.

Basic Results: Significant Improvement on Evidence and Tradeoffs

Student work revealed higher scores over time on measures of using evidence and making tradeoffs. The results from comparing the basic Evidence and Tradeoffs scores, including both Using Evidence and Using Evidence to Make Tradeoffs, portray progress for both classes. Note that throughout this time both classes were using SEPUP activities and so would be expected to improve on Evidence and Tradeoffs measures. It appears this was true, in addition to the CM group showing marked improvement on the delayed posttest. Their

average scores (on a 0-4 scale, 4 is high) are shown in Table 3.

Table 3. Average Evidence and Tradeoffs scores.

The units are from the Evidence and Tradeoffs 0-4 scale. P values are shown in parentheses.

Class	Pretest	Posttest	Delayed Posttest
Control	1.87	2.35	2.92
Experimental	1.52	1.98	3.00
Difference	.35 (<.001)	.37 (<.05)	-.08 (>.05)

Table 4. Average Evidence and Tradeoffs Gains.

The units are from the Evidence and Tradeoffs 0-4 scale. P values are shown in parentheses.

Class	Pre to Posttest	Post to Delayed Posttest	Total: Pre to Delayed Posttest
Control	.48 (<.001)	.57 (<.001)	1.05 (<.001)
Experimental	.46 (<.001)	1.02 (<.001)	1.48 (<.001)

All of these differences were significant, except for the difference between the control and experimental class on the delayed posttest (p values shown in Table 3). The improvement for the experimental group was numerically larger than for the control group; the gain scores are shown in Table 4. The experimental group shows the most improvement, not immediately after using *Convince Me*, but on the delayed posttest. A sample of work from one of the students (alias "Jamie") who improved over time follows. Jamie's answer on the pretest received a low score because it did not employ scientific evidence according to the question.

> **Question:** Should we allow human cloning for research or medicine? Should any cloning experiments be done? Give evidence from the articles to support your view. Think in terms of both advantages and disadvantages.
>
> **Response:** *Jamie's answer does not include scientific evidence from the activity and received low scores:*
>
> I think that there should not be human cloning because if you clone a person you would have same DNA and everything and if the identical clone go and do something bad, like killing somebody, and when the police go catch the person, they might find the wrong person and they can not say anything because the clone and the person have everything the same. In a way I think that allowing human cloning is good because the clone could be much healthier and everything.
>
> **Score:** Using Evidence: 1 Using Evidence to Make Tradeoffs: 1

Jamie was in the control class and did not take part in the decision-making instructional treatment. Three months and eight items later, Jamie used more evidence on the posttest:

> **Question:** Pretend you are in charge of transporting things by car two hours away. You need to pack the fragile items carefully so they are not harmed. You may pick from the following...etc.
>
> **Response:** I think I would use a card board box because it can hold the skull perfectly and that we are driving a car so we will not be scared that the skull will fall on the water and the recyclable and it will break the skull. If we use newspaper the skull will roll around the car and will break. The plastic bucket can not be recycled and the can make the skull float around so it might break. Water can make the dirty stuff away, but it may be important.
>
> **Score:** Using Evidence: 3

After another month and 6 questions, Jamie's delayed posttest answer received the highest marks. Recall that the score only reflects the criteria described in the scoring guide. Jamie's answer is not "well written," but shows "higher-level reasoning" because of the way the evidence is weighed on both sides of the issue:

> **Question:** (see Table 2, #19,20)
>
> **Response:** If I was the principal of the school I would use the plexiglass because the plexiglass it can not be break easily and if they are having so much problem and they still use the same kind of material I think they should change it. The plexiglass cost 25% more, but think of the safety of the student and not replacing it so much time. If you replace it a lot it will cost you more and it might hurt a student and for putting it up you will be interuppting the student and the worker will need you to pay for them then you will use more because every time they put it up that is hundreds of dollar already, with plexiglass it will not break easily and there will not be a lot of interruptness and hurt and you don't have to pay as much time to the workers as the glass window.
>
> **Score:** Using Evidence: 4 Using Evidence to Make Tradeoffs: 4

Rasch Results: No Progess on Evidence and Tradeoffs

The Rasch estimates of students' abilities did not indicate improvement for the two classes. The ability estimate represents the Rasch model's prediction of each student's ability on Evidence and Tradeoffs. The units are expressed in logits

$(\log[\pi/(1-\pi)]$ where π =response probability according to the model). The Rasch estimates of students' abilities that were anchored to a previous analysis as described above, showed that the classes' average estimates were worse on the final ten questions than on the first ten questions (see Table 5). However, the changes were not significant. The standard deviation on the preexam was 1.44 and on the postexam was .79 putting the two scores within reach of each other.

Table 5. Anchored Rasch Estimates of Evidence and Tradeoffs

Units are expressed in logits.

Class	Preexam (1-10)	Postexam (11-20)
Control	-.76	-1.66
Experimental	-.70	-1.55

Also, when the logit estimates are plotted onto the map of performance levels, the preexam and postexam levels are both at 2 (see Figure 2). This indicates that when item difficulty is controlled, neither class improves, but both remain at the same level. (Level 2 is similar to a score of 2 on the Evidence and Tradeoffs scoring guide.)

Figure 2. Evidence and Tradeoffs Levels

EXP=Experimental Class
CRL=Control Class

Logits	Average Preexam	Average Postexam	Levels of Performance
3.2			**Level 3** Provides relevant and accurate evidence for each claim or for at least two perspectives on the issue.
1.6			
0	EXP CRL	EXP CRL	**Level 2** Some reasons offered but evidence is incomplete or only one perspective is provided.
-1.6			

The two class's average pre and post responses were not qualitatively different according to the chart. Both classes began with an average ability level of 2, meaning that they offered reasons for their decision, but did not provide sufficient relevant evidence. Both classes ended up with an average ability level of 2 as well.

Discussion

The results demonstrated that some students from both classes became more sophisticated at answering questions that required them to use evidence and make tradeoffs. The basic analysis showed that, as expected, both groups using SEPUP activities scored higher on Evidence and Tradeoffs over time. The group using CM had significantly better posttest scores than the control group, indicating that CM was a useful tool for helping students learn to make evidence-based decisions. The experimental group's average gain from the pretest to the delayed posttest was 1.48, while the control group's was 1.08--both representing qualitative differences of at least one score level. Interestingly, the experimental group's greatest gain was between the posttest and delayed posttest. This result might imply that students benefited from integrating their experience with CM with additional SEPUP activities and evaluations. However, the Rasch estimates did not indicate improvement for the average of either class.

There are two reasons why the Rasch analysis might not have showed improvement. One possibility is that with moderate improvement (as shown in the basic analysis), one could not expect significance without a larger sample of students. A second possible reason is that the Rasch analysis compared two long time periods. The preexam took place over three weeks and there was improvement during that time; the postexam also consisted of several questions over three weeks during which there was improvement (e.g., shown by the basic analysis gains). The improvement during the lengthy preexam and postexam could mask the general improvement overall from preexam to postexam. The Rasch analysis had practical constraints preventing the administration of multiple questions on the same day: the analysis required several items in order to model the results; the items were complex questions, often associated with a laboratory activity, and took at least half an hour to answer. Thus, it was not possible to give more than one question per day.

The results suggest that *Convince Me* helped students use evidence and weigh tradeoffs in their argument. From the researcher's daily experience in the classroom, it appeared that students learned that to build an argument in CM, one was **obligated** to include evidence and not just opinions. The correlational data support, but do not prove or negate, the original hypothesis that CM's principles of coherence would enhance students' use of evidence. Still, other aspects of interacting with CM could have been beneficial beyond the principles of coherence. For instance, CM scaffolded students in connecting supporting statements and linking conflicting statements into a web. This type of reasoning experience might have been helpful when answering the Evidence and Tradeoffs questions. The act of checking one's decision after receiving feedback from the computer, then revising it, may have also been useful for the students. Further studies are necessary before a full assessment of CM's assets and flaws can be made. Previous studies have targeted parts of the program that assist learning in particular ways (e.g., Ranney et al., 1994). Current analyses of students' use of *Convince Me* will provide further evidence about students' development of decision-making skills. One reason

this research is essential is that developing ways of enhancing these decision-making skills is vital for educating students as critically-thinking citizens, as noted in many national proposals for science education reform (e.g., NRC, 1996).

References

Adams, R.J., & Khoo, S-T. (1993). *Quest: The Interactive Test Analysis System.* Hawthorn, VIC: Australian Council for Educational Research.

CTB (1997). *TerraNova.* CTB/McGraw-Hill.

Bell, P. & Linn, M.C. (1997). Scientific arguments as learning artifacts: Designing for learning on the web. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47, 149-174.*

National Research Council (1996). *National Science Education Standards.* Washington, D.C.: National Academy Press.

Ranney, M. & Schank, P. (1998). Toward an integration of the social and the scientific: Observing, modeling and promoting the explanatory coherence of reasoning. In S. Read & L. Miller (eds.)*Connectionist and PDP models of social reasoning.* Hillsdale, NJ: Lawrence Erlbaum.

Ranney, M., Schank, P., Hoadley, C. &Neff, J. (1994). 'I know one when I see one:' How (much) do hypotheses differ from evidence? In: *Proceedings of the Fifth Annual American Society for Information Science Workshop on Classification Research,* 139-156.

Roberts, L., Wilson, M. & Draney, K. (1997). The SEPUP assessment system: An overview. *BEAR Report Series SA-91-1.* Berkeley, CA: University of California.

Schank, P. K. (1995). *Computational Tools for Modeling and Aiding Reasoning: Assessing and Applying the Theory of Explanatory Coherence.* Unpublished Doctoral Dissertation, University of California at Berkeley.

Schank, P., Ranney, M. & Hoadley, C. (1996). *Convince Me* [Revised computer program (on CD, etc.) and manual]. In: J.R. Jungck, V. Vaughan, J.N. Calley, N.S. Peterson, P. Soderberg, & J. Stewart, (Eds.), The 1996-1997 BioQUEST Library (fourth edition). College Park, MD: Academic Software Development Group, University of Maryland.

Siegel, M.A. (1997). Developing decision-making skills with *Convince Me.* In: Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, 1049. Mahwah, NJ: Erlbaum.

Sloane, K., Wilson, M. & Samson, S. (1996). Designing an embedded assessment system: From principles to practice. *BEAR Report Series-96-1.* Berkeley, CA: University of California.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12,* 435-502.

Wilson, M., Sloane, K., Roberts, L. & Henke, R. (1995). SEPUP Course I, *Issues, Evidence and You:* Achievement evidence from the pilot implementation. *BEAR Report Series-SA-95-2.* Berkeley, CA: University of California.

Wright, B.D. & Masters, G.N. (1982). *Rating Scale Analysis: Rasch Measurement.* Chicago: Mesa Press.

Acknowledgements

I would like to thank Michael Ranney, Mark Wilson, Christine Diehl, and the Reasoning Group for comments on earlier drafts of this paper. I also extend much gratitude to the teacher and students who participated in my study. This research was supported by a traineeship from the National Science Foundation, Reforming Education through Science and Design, and other support from the University of California at Berkeley.