Jl. of Interactive Learning Research (2003) **14**(3), 263-283

Investigation of the "*Convince Me*" Computer Environment as a Tool for Critical Argumentation about Public Policy Issues

STEPHEN ADAMS
California State University, Long Beach, USA
sadams2@csulb.edu

The *Convince Me* computer environment supports critical thinking by allowing users to create and evaluate computer-based representations of arguments. This study investigates theoretical and design considerations pertinent to using *Convince Me* as an educational tool to support reasoning about public policy issues. Among computer environments designed to support argumentation, *Convince Me* is unique in that it computes a measure of an argument's coherence and presents this information to users as feedback. This measure is based on the *ECHO* computational model, a connectionist implementation of the Theory of Explanatory Coherence. The study seeks to better understand this coherent argumentation measure by comparing it to other measures, including a measure of the stability of one's views and the number of statements in an argument. Ten 17-year-old students and one scientist used *Convince Me* to create arguments about policies designed to ameliorate global warming; they also participated in pre- and post-intervention surveys and interviews. Positive correlations were found among the coherent argumentation measure, measures of stability, and number of statements in an argument, and these findings raise considerations for designing educational activities with *Convince Me*. A debriefing interview's results illustrate further considerations, including the role of the user's stance towards the software.

The process of argumentation, including evaluating evidence related to differing assertions, is central to critical thinking (Kuhn, 1991, 1992). This study investigates reasoning about public policy issues using *Convince Me*, a computer environment that supports argumentation (Diehl, 2001; Diehl, Ranney, & Shank, 2001; Schank, Ranney, & Hoadley, 1999; Schank, Ranney, Hoadley, Diehl, & Neff, 1994; Siegel & Ranney, in press; Weidner, Ranney, & Steinbach, 1998). *Convince Me* is one of a variety of computer environments that support users in creating and evaluating arguments. Such environments include Belvedere (Suthers, Connelly, Lesgold, Paolucci, Toth, Toth, & Weiner, 2001; Suthers & Weiner, 1995; Suthers, Weiner, Connelly, & Paolucci, 1995), CSILE (Scardamalia & Bereiter, 1991), and Sensemaker (Bell, 1997, 1998; Bell & Linn, 2000). The *Convince Me* argument construction software is unique in that it computes a value that is derived from comparing a user's argument and belief evaluations to a model of coherent reasoning. These "Model's Fit" values represent a construct central to *Convince Me* and are taken as a kind of desideratum of good reasoning. They are derived using a connectionist computer model called *ECHO*, which is a computational implementation of the Theory of Explanatory Coherence developed by Thagard (1989). The present study seeks to contextualize the environment's Model's Fit measure by comparing it to other measures, including measures of the stability of participants' views. For further experimental information, participants were asked in a debriefing interview to discuss how well they thought their *Convince Me* arguments reflected their thinking. The study uses policies proposed to ameliorate global warming as a context, since global warming raises highly important public policy questions for citizens.

THEORETICAL BACKGROUND

Previous research investigating argumentation in the context of societal issues has included topics including crime, education, and unemployment (Kuhn, 1991, 1992), as well as problem solving in international relations (Voss, Lawrence, & Engel, 1991). Belvedere, a computer environment that supports argumentation, was studied initially as a tool for 12- to 15-year-olds in areas related to science and public policy issues (Suthers, Weiner, Connelly, & Paolucci, 1995) and more recently as a tool to teach college students problem-solving in economics (Cho & Jonassen, in press). Toulmin's (1958) model of argument, delineating components of arguments including data, warrants, backings, and conclusions, has been influential in much of this work.

Designers of computer environments to support argumentation must make assumptions about what constitutes a good argument and assumptions about how the process of constructing such arguments may best be supported. For example, the design of the initial implementation of Belvedere involved the creation of graphical icons for various argument components inspired by categories of Toulmin's (1958) model of argument. Although it has some similarities to other computer environments to support argumentation, the *Convince Me* computer environment is unique in its strong ties to a philosophical and theoretical tradition that better arguments and better thinking involve more coherent explanations. Given that multiple theoretical choices are possible when designing a computer environment to support argumentation (e.g., using Toulmin's approach or using explanatory coherence), this study is oriented toward understanding more about the advantages and limitations of *Convince Me*'s design.

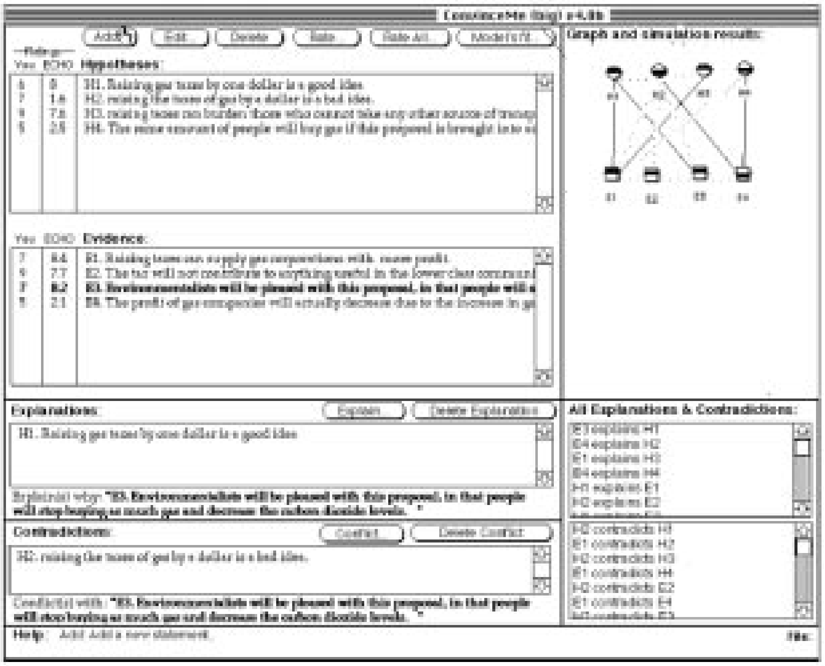
Perspectives valuing coherence as emblematic of good reasoning are seen in the literatures of political science, public opinion research, and cognitive science. In political science, an influential paper by Converse (1964) posited that holding a coherent political ideology is a sign of greater sophistication. In Converse's view, an ideology is coherent if a person's position on one issue predicts his or her position on another issue.

Similarly, in the area of public opinion research, Yankelovich used coherence as an indicator of a better opinion. He proposed two formal criteria for the quality of opinion about policy issues: consistency (whether the opinion was consistent with one's other beliefs) and volatility (whether the opinion was firmly held or changes) (Yankelovich, 1991, p. 24). Yankelovich, Skelly, and White (1981) developed a measure designed to gauge, relatively quickly, the latter of these criteria, volatility (or stability). The measure, which Yankelovich colorfully named the "Mushiness Index," consisted of a set of four questions. Three questions related to hypothesized sources of stability: whether a person had a personal stake in an issue, had more information about an issue, or has had discussions with others about the issue. A fourth question simply asked respondents to estimate the likelihood that they will change their minds. Yankelovich et al. (1981) found that the index was relatively good at predicting whether persons would change their positions about a policy issue. The present study uses Yankelovich's Mushiness Index, but renames it the "Stability Index" for greater clarity, since higher values on the index correspond to higher stability.

In the area of cognitive science, techniques employing connectionist models also provide a way to gauge the coherence of a set of propositions. Thagard (1989) developed a Theory of Explanatory Coherence, embodied in *ECHO*, a computational model of reasoning (Ranney & Thagard, 1988). The Theory of Explanatory Coherence is based on a set of principles with psychological interpretations such as: (a) all other things being equal, a statement is more believable if it is supported by evidence, and less believable if it is contradicted by strong alternatives; (b) a statement is more believable if it is supported by plausible beliefs and less believable if it is supported by implausible beliefs (Ranney & Schank, 1998). *ECHO* has been used to model scientific controversies in many areas, including the history of science (Thagard, 1989) and physics reasoning (Ranney & Thagard, 1988). In these studies, coherent argumentation according to *ECHO* was employed to evaluate reasoning. Further, Ranney and Schank (1998) utilize *ECHO* to draw parallels between scientific and social reasoning.

Convince Me allows users to create and evaluate representations of their own thinking. Among computer environments that support the creation of arguments, *Convince Me* is unique as it provides feedback based on the *ECHO* computational model. This feedback serves as a prompt for users to reassess their arguments. Figure 1 shows the interface of the *Convince Me* program.

The software provides a way for a user to enter a set of statements and categorize them as either hypotheses or evidence (Figure 1, upper left quadrant). The user then creates a set of links among the statements, indicating which statements explain or contradict others. The program displays these relationships graphically (Figure 1, upper right quadrant). In addition, the user enters a set of Believability ratings indicating how much he or she believes each individual statement (using a scale where "1" is low and "9" is high). The program does not understand the meaning of the statements, but when the *ECHO* model is run, it generates activation values that may be thought of as its own evaluations of the plausibility of the statements, based on the structure of the argument and *ECHO*'s rules. The program displays these model-generated values adjacent to the user's Believability ratings



Ratings

You	ECHO	Hypotheses:
6	8	H1. Raising gas taxes by one dollar is a good idea.
7	1.6	H2. raising the taxes of gas by a dollar is a bad idea.
8	7.8	H3. raising taxes can burden those who cannot take any other source of transp
5	2.5	H4. The same amount of people will buy gas if this proposal is brought in to e

You	ECHO	Evidence:
7	8.4	E1. Raising taxes can supply gas corporations with more profit.
9	7.7	E2. The tax will not contribute to anything useful in the lower class communi
7	8.2	**E3. Environmentalists will be pleased with this proposal, in that people will s**
5	2.1	E4. The profit of gas companies will actually decrease due to the increase in ga

Explanations:

H1. Raising gas taxes by one dollar is a good idea.

Explained why: **"E3. Environmentalists will be pleased with this proposal, in that people will stop buying as much gas and decrease the carbon dioxide levels. "**

Contradictions:

H2. raising the taxes of gas by a dollar is a bad idea.

Conflicted with: **"E3. Environmentalists will be pleased with this proposal, in that people will stop buying as much gas and decrease the carbon dioxide levels. "**

All Explanations & Contradictions:

E1 explains H1
E4 explains H2
E1 explains H3
E4 explains H4
H1 explains E1
H2 explains E2

H2 contradicts H1
E1 contradicts H2
H2 contradicts H3
E1 contradicts H4
H2 contradicts E2
E1 contradicts E4

Help: Add: Add a new statement.

file:

Figure 1. The *Convince Me* interface

(Figure 1, upper left quadrant). For example, a statement categorized as evidence is more likely to be highly activated than one categorized as a hypothesis; a statement is also more likely to be highly activated if it explains more statements than another. Finally, the program can report a Model's Fit value, which represents the correlation between the user's Believability ratings and *ECHO*'s model-generated activation values. The program reports these Model's Fit values to users, and it also notes those statements for which the user's believability ratings differ most sharply from the values produced by the model. With this feedback, users may revise the structure of their argument. They may also adjust their ratings of the believability of any statement in an argument.

An assumption of the study is to consider these Model's Fit values to be indicative of coherent argumentation, and this is supported by both theoretical reasons and previous experimental findings. As noted earlier, from a theoretical perspective, Model's Fit values are derived from *ECHO*, which in turn is based on the Theory of Explanatory Coherence. From an experimental perspective, *ECHO* has been shown to be successful at modeling students' reasoning and believability ratings (Schank & Ranney, 1991). Nonetheless, an important caveat to note is the assumption that a *Convince Me* argument reflects what the participant really knows. Some users may not articulate their beliefs well. For this reason, although the Model's Fit values will be referred to as a proxy of coherent argumentation, it should be emphasized that the measure is more precisely described as the correlation between the user's Believability ratings and the corresponding values for each statement generated by the *ECHO* model.

A body of studies with *Convince Me* have been conducted (Diehl, 2001; Diehl, Ranney, & Schank, 2001; Ranney & Schank, 1995; Ranney, Schank, Hoadley, & Neff, 1996; Schank, 1995; Schank, Hoadley, Dougery, Neff, & Ranney, 1993; Schank & Ranney, 1991, 1992,1993; Schank, Ranney, Hoadley, Diehl, & Neff, 1994; Siegel, 1999; Siegel & Ranney, in press). Ranney, Schank, & Diehl (1995) provide an overview of the rationale for *Convince Me* and synthesize early studies with the program.

Schank (1995) compared groups of subjects who created arguments either using *Convince Me* or using paper-and-pencil. Compared to the paper-and-pencil group, users of *Convince Me* created arguments with higher Model's Fit values; the *Convince Me* group also made more changes to the structure of their arguments. Also, the increase in Model's Fit ratings was generally attributed not to users modifying their belief ratings to match *ECHO*'s, but to further explicating their arguments. Diehl (2001) compared four groups of students using *Convince Me* (with a fifth group serving as a control). The groups differed according to whether they received feedback from the *ECHO* model and whether they worked either in pairs or individually. Their arguments were assessed both by their Model's Fit ratings and by a separate argument measure. Receiving feedback from the model improved participants' Model's Fit values, while there was no effect on Model's Fit values derived from working in pairs versus working individually. On the other hand, either receiving feedback from the model or working in pairs (or both) yielded better performance on the argumentation measure.

In other words, in previous studies, use of *Convince Me* has led to better elaborated arguments, higher Model's Fit values, and better scores on other measures of argumentation. The subject areas used in previous studies with the program have included topics related to scientific issues, decision-making, and public policy. This study explores some design assumptions of *Convince Me* that that have not been investigated in prior studies and are especially pertinent to its application as a support for reasoning about public policy issues. The study investigates two primary hypotheses. Hypothesis P-1 is that stability and coherence may be related, that is, that Stability Index scores may be correlated positively with Model's Fit values. From a theoretical perspective, this hypothesis is of interest because, if supported, it would show an alignment between two rather disparate measures derived from different theoretical traditions in cognitive science (*Convince Me*'s Model's Fit values) and public opinion research (the Stability Index). Also, because *Convince Me*'s Model's Fit value is used as a measure of a good argument, identifying correlates and possible influences on this measure is of interest to the task of designing educational activities with the program. The study also investigates the relationship between the number of statements in *Convince Me* arguments and their Model's Fit values. Hypothesis P-2 is that Model's Fit values may be negatively correlated with the number of statements in corresponding *Convince Me* arguments. Specifically, it was hypothesized that the increased complexity of *Convince Me* arguments that involve more statements would make it more difficult to create a coherent argument. On the other hand, the converse hypothesis is also plausible: one could hypothesize that arguments with more statements would reflect a greater knowledge base about the topic and would therefore be likely to be more coherent. Again, from an educational design perspective, exploring this hypothesis is useful in identifying factors related to the Model's Fit values.

In addition, the study investigates three hypotheses which provide supporting information for interpreting the primary hypothesis. Hypothesis S-1 is that participants in the study (both high school students and a postdoctoral researcher) would be able to use *Convince Me* to produce well-formed arguments about policy questions, yielding moderately positive Model's Fit values (again, each Model's Fit value represents the correlation between the user's Believability Ratings and the activations from the *ECHO* model). Although the intervention was not designed to promote particular policies, Hypothesis S-2 is that participants' support for policies would increase as a result of learning more about global warming. Hypothesis S-3 is that the study would support the findings of Yankelovich et al. (1981) that the Stability Index predicts changes in positions; that is, higher scores on the Stability Index would correspond to fewer changes in positions. A further research question was more open-ended: by asking participants to discuss their uses of *Convince Me* after they have created arguments with it, the study probes for their views of the software and possible influences on their use of it. Identifying such information is of interest to the task of designing educational activities with *Convince Me*, or related programs to support argumentation.

METHODS

Participants

A total of 11 participants (10 high school students and one scientist) took part in two experimental sessions totaling about 3.75 hours. The students were all seventeen years old and seniors from a single San Francisco Bay Area high school. To increase representativeness, students were drawn from science classes having students of mixed ability levels. The students had all taken three years of high school science classes, but their GPAs in these classes ranged from 2.6 to 4.0. Of the ten students, four were male and six were female. Two students were African-American, one student was Asian-American, and seven were Caucasian. For additional diversity, in terms of previous science knowledge, a scientist was also included. He was a 34-year-old male postdoctoral scholar studying in the field of climate change at a major research university. Students were paid $5.75 per hour, while the postdoctoral researcher was paid $12.50 per hour. Because the goal of the study was not to serve as a large-scale summative assessment but rather to serve as a formative investigation of using *Convince Me* as a support for reasoning about science-based policy issues, the sample size was judged to be sufficient for these initial purposes. (Note that in the Results, the names are fictitious but preserve gender.)

Materials

The materials included the *Convince Me* software and instruction manual, a Policy Questionnaire, and a Stability Index questionnaire based on five of the policies regarding global warming. The *Convince Me* software (Schank, Ranney, & Hoadley, 1999) was used together with a set of instructions adapted from Schank's (1995) instruction manual. The Policy Questionnaire, which was adapted from one developed by Public Agenda (Doble & Johnson, 1990; Doble, Richardson, & Danks, 1990), asked participants to rate their support for 15 policies that have been proposed to ameliorate global warming. For each policy, participants were asked to choose one of the following options:

1. Do it immediately

2. Phase it in gradually, over the next 10 years or so
3. Don't do it, no matter what
4. Not sure.

The format of the questionnaire explicitly included a trade-off for each proposed policy, as illustrated in the following two policies:

- Raise the gasoline tax by $1.00 a gallon *even if* that would burden truckers and others who need their cars for work.
- "Fee-rebate" system. Charge people who buy cars with poor gas mileage an additional fee and use the money to give rebates to people who buy cars with good gas mileage. People who buy cars with good gas mileage would get rebates of up to $1,000 *even if* people who buy cars with poor gas mileage would be charged fees of up to $1,000.

As discussed below in further detail, participants were later interviewed about these two policies and asked to create arguments using *Convince Me* that described their views about them. In addition, participants were interviewed about three other policies from the Policy Questionnaire: increasing the use of nuclear power, fertilizing the oceans with iron to stimulate phytoplankton (and thereby sequestering carbon dioxide), and another automobile-related policy that would require raising the fuel efficiency (mpg) requirements for automobiles. In this way, the study incorporated a variety of different types of policies, involving regulations, incentives, and technological and geoenginnering approaches. Policies affecting automobiles were generally emphasized both because of the expectation that they would be the most tangible to participants and because of the substantial contribution to global greenhouse gas output attributed to the use of automobiles in the United States.

The Stability Index questionnaire, designed to probe the stability of participants' views, consists of a set of four questions as shown in Table 1. Participants were given the Stability Index questionnaire for each of the five policies included in the interview. For each policy, participants were asked to rate their positions on each question on a scale of 1 to 6. The overall score is the sum of the scores on the individual questions, yielding a scale ranging from 4 to 24.

In the Debriefing Interview, participants were asked: "What did you think about the *Convince Me* program?" and "How well do you think your *Convince Me* argument represents your knowledge and beliefs?" Participants were asked follow-up questions on a case-by-case basis.

Procedure

There were two experimental sessions within five days, which lasted about 1 3/4 and 2 hours, respectively. Figure 2 gives an overview of the experimental activities.

In the first session, participants were introduced to the purpose of the

Table 1. The Questions of the Stability Index

a. On a scale of 1 to 6, where 1 means that the issue affects you personally *very little* and 6 means that you really feel *deeply involved* in this issue, where would you place yourself?

b. On some issues people feel that they really have all the information that they need in order to form a strong opinion on that issue, while on other issues they would like to get additional information before solidifying their opinion. On a scale of 1 to 6, where 1 means that you feel you definitely need *more* information on the issue and 6 means that you do *not* feel you need to have any more information on the issue, where would you place yourself?

c. On a scale of 1 to 6, where 1 means that you and your friends and family *rarely, if ever*, discuss the issue and 6 means that you and your friends and family discuss it relatively *often*, where would you place yourself?

d. People have told us that on some issues they come to a conclusion and they stick with that position, no matter what. On other issues, however, they may take a position, but they know that they could change their minds pretty easily. On a scale of 1 to 6, where 1 means that you could change your mind *very easily* on this issue and 6 means that you are likely to *stick with your position* no matter what, where would you place yourself?

Note: From Yankelovich et al. (1981)

study and given a pretest Policy Questionnaire (for 15 policies) and a pretest Stability Index Questionnaire (for 5 policies). Then, in preparation for the second session, they participated in interviews and a briefing concerning the science and uncertainties of global warming, reports of global warming from the media, and some policy options proposed to ameliorate global warming. The briefing materials incorporated a non-partisan booklet about global warming written for lay persons by researchers at Carnegie-Mellon University (Morgan & Smuts, 1994). (For content analyses of these interviews, see Adams [1999, 2001]). In the second session, participants were given a briefing about the *Convince Me* software and provided instructions on how to use it. They were then interviewed and subsequently created *Convince Me* arguments about the gasoline tax and fee-rebate policies. For both policies, participants also took the Stability Index questionnaire immediately after completing their *Convince Me* argument for that policy. They also took the full Policy Questionnaire a second time after completing their *Convince Me* arguments. Finally,

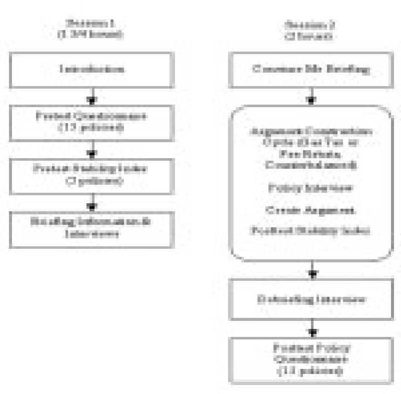


Figure 2.
Experimental Procedureprocedure

participants were given a Debriefing Interview regarding their experiences and reflections about the experimental session. The interviews were audiotaped and transcribed.

Administering the Policy Questionnaire at the beginning and at the end of the study provided a way to check for any changes in participants' positions. For two of the policies—the gasoline tax and the fee-rebate system—participants also created *Convince Me* arguments on the computer. These policies were selected since they both involved the common theme of the automobile, while they presumably differed in the degree to which participants were likely to have firm opinions about them. Specifically, it was expected that participants would have firmer opinions about the gasoline tax proposal than the fee-rebate proposal. Whereas the gasoline tax is a more conventional proposal and such taxes tend to be unpopular in the United States, the fee-rebate proposal is less conventional and it was expected that lay persons would be less likely to have formed opinions about it. The order of these two *Convince Me* exercises was counterbalanced across the participants.

As an indicator of the stability of participants' support for the five policies, the Stability Index was administered to participants at the beginning of the first experimental session. It was hypothesized that the experimental interventions, including the briefing and interviews about the policies, could possibly increase participants' overall knowledge about the issues. Also, by providing a setting for participants to think through the implications of adopting or not adopting the policies, the intervention could possibly prompt participants to change their positions. Therefore, the Stability Index questions relating to each *Convince Me* argument were administered a second time after the interviews. This second Stability Index questionnaire was conducted immediately before the Model's Fit value was taken so that the two measures would reflect students' cognitive states following the interview and *Convince Me* exercise. In other words, the second Stability Index score provides a temporally critical comparison to the Model's Fit value. On the other hand, the first Stability Index score, given before the interviews, was taken to determine whether the Stability Index inversely correlates with changes in the level of support for a policy.

RESULTS

Table 2 summarizes the experimental findings. The results supported all hypotheses except Hypothesis P-2, for which the converse hypothesis was better supported.

The discussion of the results is organized in four main parts. Parts 1a and 1b discuss the supporting hypotheses, Part 2 discusses the primary hypotheses, and Part 3 discusses the debriefing interviews.

Table 2. Summary of Experimental Findings

Primary Hypotheses	**Supported?**
P-1. Stability Index values will be positively correlated with Model’s Fit values	Yes (*r* = .48)
P-2. Model’s Fit values will be negatively correlated with the number of statements in an argument	No (*r* = .34)
Supporting Hypotheses	
S-1. Model’s Fit values will be moderately positive	Yes (*r* = .57)
S-2. Support for policies will tend to increase	Yes (see below)
S-3. Higher values on the Stability Index will correspond to fewer position changes	Yes (see below)

1a. *Convince Me* Arguments

Hypothesis S-1, that study participants would produce *Convince Me* arguments with moderately positive Model’s Fit values, was supported. Descriptive statistics indicate that participants’ *Convince Me* arguments were well-formed overall. The average Model’s Fit value (r = .57) was similar to the average Model’s Fit value in Schank’s (1995) study, which used college students (r = .53). Overall, Model’s Fit values for individual arguments ranged from -.32 to 0.98. The scientist’s Model’s Fit values were toward the higher end of the range: 0.87 for each problem. If the scientist’s Model’s Fit values are dropped, the average Model’s Fit value decreases only slightly (r = .54). One Model’s Fit value, for a student’s gasoline tax problem, was anomalously low (-0.32); it will be described further in the “Debriefing Interview” section. Model’s Fit values were somewhat higher for the fee-rebate problem (r = .62) than the gasoline tax problem (r = .52). However, this difference was not significant: $t(20) = 0.65$, $p < 0.53$. Table 3 provides descriptive statistics of participants’ *Convince Me* arguments.

1b. Changes in Support for Policies

Hypothesis S-2 was supported, in that participants’ support for policies increased somewhat from pretest to posttest. Support for policies was coded numerically from the Policy Questionnaire, in which opposing a policy was coded as -1, supporting a policy for implementation either immediately or to be phased in was coded as +1, and being not sure was coded as 0. On the pretest, the mean overall support for the five policies was -0.23 (somewhat opposed), with a standard deviation of 0.92. By the posttest, the average support increased to +0.36 (somewhat supportive), with a standard deviation of 0.95. Thus, the mean difference increased by 0.59, which is significant, $t(21) = 2.89$, $p < .01$. This is consistent with the expectation that the experimental intervention would serve to heighten participants’ knowledge about global

Table 3. Descriptive Statistics of *Convince Me* Arguments

	mean			s. d.			minimum			maximum		
	fee	tax	both	fee	tax	both	fee	tax	both	fee	tax	both
Mf[a]	.62	.52	.57	.36	.37	.36	-.11	-.32	-.32	.98	.88	.98
H[b]	4.3	5.0	4.6	2.6	3.4	3	2	2	2	10	12	12
E[c]	5.0	5.2	5.1	2.1	2.2	2.1	1	1	1	8	9	9
H+E	9.3	10.3	9.8	2.6	2.2	2.4	4	7	4	14	13	14
Ex[d]	6.5	7.3	6.9	3.1	3.1	3.1	2	3	2	13	15	15
C[e]	6.3	7.3	6.8	3.0	2.7	2.8	1	1	1	10	11	11
Ex + C	12.8	14.5	13.7	4.5	4.7	4.6	5	8	5	17	25	25

[a]Model's Fit values (i.e., belief rating-activation correlations) [b]Hypothesis statements. [c]Evidence statements. [d]Explanation links. [e]Contradiction links.

warming and their willingness to support policies to ameliorate it.

Support for the fee-rebate policy increased dramatically while support for the gasoline tax policy did not change much. On a numeric scale in which -1 denotes opposition and +1 denotes support for a policy, mean support ratings for the fee-rebate policy rose significantly, from -0.27 to 0.64, $t(10)$ = 3.19, p < 0.01. On the other hand, mean support ratings for the gasoline tax policy were -0.18 on the pretest and 0.09 on the posttest, which was not a significant increase, $t(10)$ =1.00, p < .34.

A position change was defined as shifting between one of the following three categories: supporting a policy, not supporting a policy, or not being sure. Experimental findings were consistent with Hypothesis S-3 in that higher (more stable) ratings on the Stability Index corresponded to fewer position changes. Overall, the average pretest Stability Index score for all five policies was 11.6 (S. D. = 4.6) on a scale in which 4 represents the least firm beliefs and 24 represents the most firm beliefs. Because 11 participants were interviewed and took the Stability Index questionnaire for five policies, there were a total of 55 opportunities for participants to change their positions about their support for a policy. There were sixteen instances of such changes (29%). The mean pretest Stability Index score for those policies with a position change, 9.7, was significantly less firm for those for which there was no position change, 12.4, $t(53)$ = 2.01, p < .05). In other words, higher ratings on the Stability Index corresponded to fewer position changes.

2. Correlations Among Measures

Hypothesis P-1, that Model's Fit values would be positively correlated with Stability Index values, was supported for posttest Stability Index scores. The correlation between Model's Fit values and Stability Index posttest scores was 0.48 ($p < 0.05$); without the outlier corresponding to the anomalously low Model's Fit value, this increased to .55 ($p < .01$). The correlation between Model's Fit values and Stability Index posttest scores was also higher in the fee-rebate problem ($r = .63$, $p < .05$) than the gasoline tax problem ($r = .51$; $p = .11$). Although there was a positive correlation between posttest Stability Index scores and Model's Fit values, pretest Stability Index scores were not significantly correlated with Model's Fit values ($r = .21$, $p = .36$). A test of whether these correlations differed (i.e., .21 vs. .48) was not statistically significant ($z = 1.38$; $p = .17$).

Hypothesis P-2, that Model's Fit values would be negatively correlated with the number of statements, was not supported. To the contrary, the correlation between Model's Fit values and the number of statements in a *Convince Me* argument was 0.34 ($p = 0.12$); without the outlier, this rose to 0.50 ($p < .05$). The correlation between the number of statements in an argument and Stability Index posttest scores was 0.45 ($p < 0.05$); without the outlier, the statistic remained the same. A further finding was that pre-test Stability Index scores were also positively and significantly correlated with the number of statements in a *Convince Me* argument ($r = 0.56$; $p < 0.01$); the correlation remained the same without the outlier ($r = 0.56$; $p < 0.01$). This is consistent with the interpretation that participants with firmer beliefs about a policy would tend to create arguments with more statements.

The positive correlation between Model's Fit values and Stability Index ratings supports the theoretical expectation that stability and coherence are related (Hypothesis P-1). However, did Model's Fit values correlate more closely to some particular questions within the Stability Index than to others? Table 4 shows a correlation matrix involving the individual Stability Index items. (Recall that Table 1 has the full text of the questions. The sum of these four items is the overall Stability Index score.) The item most high-

Table 4. Correlation Matrix of Individual Stability Index Items[1]

	a	b	c	d	a+b+c+d	MF[a]
a	-	.54	.68	.57	.92	.30
b		-	.24	.37	.64	.18
c			-	.27	.80	.43
d				-	.66	.57
a+b+c+d					-	.48
Mf[a]						-

[1] a = personal involvement; b = amount of information; c = discussing issue with others; d= likelihood of changing mind; MF= Model's Fit

ly correlated with Model's Fit values was Question d ($r = .57$), concerning the likelihood of changing one's mind. The item least correlated with Model's Fit values was Question b ($r = .18$), which concerned having sufficient information about the issue. Given that Question d is the most central aspect of stability, relative to questions a-c, Model's Fit seems even more impressively related to the construct of the stability of one's views.

3. Debriefing Interview

Data from the debriefing interviews were analyzed, and they provide evidence for three influences on participants' construction of arguments.

1. Students may view some statements in a *Convince Me* argument as much more important than others. As previously noted above, the Model's Fit values for one participant's gasoline tax argument was anomalously low: -0.32. The student's *ECHO* representation had six statements linked to the hypothesis he believed more strongly ("raising the gasoline tax by $1.00 is not a good idea") but only three statements linked to the opposing hypothesis that he believed less strongly. It turned out the student felt that one of the three statements ("The tax would change the whole fabric of society") was more important than all the others. The student explained:

> The real reason why I lean on the side of raising the gasoline tax by one dollar is not a good idea, is because it's gonna affect the way our society is made up…and that's the biggest factor. But I still agree with arguments such as "people aren't gonna drive as much cars"…they're both true.

This outlier shows an example of a situation that can create a problem for *Convince Me*. The program gives more weight to evidential than to hypothetical statements, but does not have a way to directly represent or weight the "importance" of a hypothesis. If a user of *Convince Me* does not link multiple statements to an "important" hypothesis, *ECHO* might tend to rate that hypothesis lower than the user did.

2. Students with more stable views on an issue may generate more statements. As indicated earlier, the data indicate that Stability Index scores were correlated with the number of statements participants wrote in a *Convince Me* argument. One of the clearest examples of this pattern came from differences in how one subject responded to the gasoline tax and fee-rebate policies. Her pretest Stability Index score for the gasoline tax problem (15) was nearly twice as firm as her score on that test for the fee-rebate problem (8). Similarly, her *Convince Me* argument for the gasoline tax problem contained twice as many statements (8) as her argument for the fee-rebate policy (4).

In the *Convince Me* debriefing interview, she said that she liked using *Convince Me* for the gasoline tax problem: "The one about gasoline, that was fun." She indicated that it was more fun than the fee-rebate problem because it was more relevant to her: "I'm not buying a car right now…so [(the fee-rebate policy]) didn't really apply to me. But gas applies to me, because I drive a car, and I drive my mom's car a lot, and my grandmother's car, so it really applies to me."

As discussed earlier, one of the questions of the Stability Index specifically checks for this kind of perceived personal relevance. (See the first question of Table 1.) Personal relevance may contribute to a subject's motivation and interest in creating *Convince Me* arguments.

3. Differing stances of users toward *Convince Me* can influence the use of the program. Participants differed in their views of using *Convince Me*, as illustrated by the contrasting responses of students. Comments of the student Marie reflect the view that *Convince Me* is useful for forming an opinion about a policy:

> That was…a good learning tool…I got more into detail and I was questioning my ideas about what kinds of businesses would need low mileage trucks and cars, and the awards for buying high mileage cars…When I did this question [(on the pretest questionnaire],), I was more indecisive. I didn't really have a point of view on the argument, and (*Convince Me*) helped me to get one, whether it was good or bad…I didn't really know if that's a good or a bad idea, but now I've kind of decided it's overall a good idea.

Consistent with her explanation that she formed a clearer opinion on the policy, her Stability Index score increased from 7 on the pretest to 11 on the posttest. When asked to compare the experience of being interviewed with using *Convince Me*, she indicated that she preferred *Convince Me*: "I think this is better, at the computer, because it, you get more time to think about things and you can make more of a strong link between different arguments, and it's a lot more concrete." On the other hand, the student Howard, while acknowledging that using the program helped him clarify his thoughts, indicated that he was critical of using it: "What was the point of it? I guess the point was sort of to clarify in my mind what I thought about these things? I think it helped about as much as writing down my own thoughts would. I don't think it was really necessary." He viewed himself as having an adversarial relationship with the computer: "That felt like I was, I should've been trying to please the computer, or trying to trick the computer. Well, for me it sets up an adversarial relationship between how stupid the computer is and how much smarter I am than it. I just don't want to play chess against it."

Just before he finished it, Howard's *Convince Me* argument for the fee-rebate problem contained three statements linked with a hypothesis supporting the fee-rebate policy but only two statements linked with the hypothesis opposing the policy. Then he added a flippant hypothesis: "Oodles of green poodles jump der strudle." He linked this statement to the hypothesis opposing the fee-rebate proposal—the side that had one fewer statement. When asked why he included this statement, he indicated he was "balancing" the argument. In effect, by adding one more statement to the argument with the fewer hypotheses, he improved the alignment between his belief ratings and how *Convince Me*'s model would evaluate the believability of the statements. Previous studies have found that most participants' attitudes towards *Convince Me* have been positive overall (Schank, 1995; Siegel, 1999; Diehl, 2001). Even if that is the case, the contrasting cases cited above illustrate the dramatic role a user's stance towards the software can play in the nature of the educational experiences derived from using it.

DISCUSSION

The *Convince Me* environment places special emphasis on explanatory coherence as a theoretical construct. While there are ample a *theoretical* grounds for incorporating explanatory coherence as a construct in the design of a computer environment to support argumentation about public policy issues, the study identifies some of the theoretical and practical considerations underlying the use of *Convince Me*'s Model's Fit metric as feedback for creating a good argument. It is noteworthy that the Model's Fit values correlated positively with the Stability Index scores (Hypothesis P-1), considering that these measures were derived from different theoretical traditions (e.g., cognitive science and public opinion research) using different methodologies (e.g., the *ECHO* connectionist computer model vs. the Stability Index questionnaire). On the one hand, more stable views about an issue could be an indicator that those views are better developed. On the other hand, more stable views could be a result of a kind of closed-mindedness or rigidity.

The finding that the Stability Index measure correlates positively with the Model's Fit values is particularly germane to the application of *Convince Me* to reasoning about public policy issues. Further, findings from the study have implications for using *Convince Me* with other subject areas, by underscoring the need to evaluate whether or not and when higher Model's Fit values may be considered more desirable. This could be explored further in subsequent research. Because of the possibility that users of the software may tend to view higher Model's Fit values as "better," it is worth emphasizing to users (a) the technical meaning of higher Model's Fit values—that they represent a higher correlation between the user's Believability ratings and *ECHO*'s estimation of the "believability" of each statement—as well as (b) caveats on interpreting these values.

Likewise, in evaluating *Convince Me* arguments, the relationship between the number of statements a user incorporates into a *Convince Me* argument and the quality of the argument it represents should also be considered. More statements could be a sign of being better informed about an issue, but could alternatively reflect only greater verbosity. Also, it is trivially easy to be coherent about an issue if one knows virtually nothing about it; good reasoning involves maintaining consistency over an appropriately large knowledge base. In addition, although the correlation between Model's Fit values and the number of statements was positive in this study (refuting Hypothesis P-2), a, a review of data from previous studies with *Convince Me* suggests this relationship seems labile in that it is not yet well understood how a variety of possible factors modulates it (Michael M. A. Ranney, personal communication, February 2, 2003). Further research with *Convince Me* (for any subject area) could explore the role of potential variables, for instance, the extent to which participants have clear evidence about the topic or must rely on hypotheses.

The data also suggest that *ECHO* models may be less robust in cases in which the relative importance of beliefs is unevenly represented in participants' *Convince Me* arguments. A way to weight the importance of statements could be incorporated into the program, or users could be encouraged to more evenly and thoroughly represent their beliefs. The debriefing interviews also illustrate ways in which the stances that participants adopt towards using the program may influence the representations they create and the educational experiences they derive from using it. This is clearly an important part of self-directed educational activities such as this one. Subsequent research could further investigate the role of students' stances towards the program.

The design problem of getting a computer program to recognize a good argument is a challenging one. From an educational standpoint, however, helping students recognize characteristics of a good argument would be a higher priority. Further, it is appropriate to judge computer representations not merely on their accuracy with respect to a model of expertise but rather in terms of the conversations they engender among their users (Roschelle, 1996). In designing educational activities with *Convince Me*, the kinds of complexities identified in this study could be turned to pedagogically useful ends. For example, to stimulate students to reflect about characteristics of good arguments, students who have created multiple arguments with *Convince Me* could be asked to develop and discuss criteria for determining whether or not and in which circumstances—they view it's explanatory coherence values as indicative of "better" or "worse" arguments. That is, the design goal would be to prompt students to evaluate not only particular arguments but also to produce generalizations about characteristics of good arguments. In this way, considerations raised by this study could serve a useful educational role not only for learning to critically evaluate arguments but also for learning to evaluate the utility and limitations of information from a computer model.

Taken as a whole, the considerations raised here caution against taking the explanatory coherence notion, the Model's Fit metric, unqualified as a measure of a good argument. However, by articulating theoretical and practical considerations pertinent to using *Convince Me* as a tool for creating arguments about public policy issues, the study provides information useful for designing educational activities with the program that capitalize on its ability to support users in articulating and evaluating arguments. In addition, the study identifies considerations pertinent to the use of *Convince Me* (e.g., factors such as the extent of the user's prior knowledge, the stability of the user's beliefs, the size of the user's argument, and the user's stance towards the software) that would also be relevant to designing other computer systems for creating and evaluating arguments.

References

Adams, S. (1999). Views of policies affecting automobiles: A comparison of high school students and specialists. *Bulletin of Science, Technology, & Society, 19*(5), 372-380.

Adams, S. (2001). Views of the uncertainties of climate change: A comparison of high school students and specialists. *Canadian Journal of Environmental Education, 6*, 58-76.

Bell, P. (1997). Using argument representations to make thinking visible for individuals and groups. In R. Hall, N. Miyake, & N. Enyedy (Eds.), *Proceedings of CSCL '97: The Second International Conference on Computer Support for Cooperative Learning* (pp. 10-19). Toronto, Ontario, Canada: University of Toronto Press.

Bell, P. (1998). *The KIE software and curriculum: Relating debate activities and conceptual change through design experiments.* Paper presented at the Annual Conference of the American Educational Research Association 1998, San Diego, CA.

Bell, P., & Linn, M. (2000). Scientific arguments as learning artifacts: Designing for learning from the web with KIE. *International Journal of Science Education, 22*(8), 797-817.

Cho, K., & Jonassen, D. (in press). The effects of argumentation scaffolds on argumentation and problem-solving. *Educational Technology: Research and Development.*

Converse, P. E. (1964). The nature of belief systems in mass publics. In D. Apter (Ed.), *Ideology and Discontent* (pp. 206-61). New York, NY: Free Press.

Diehl, C. (2001). *Computers and students students as instructional partners: The role of simulation feedback in collaborative argumentation.* Unpublished doctoral dissertation, University of California, Berkeley.

Diehl, C., Ranney, M., & Schank, P. (2001). Model-based feedback supports reflective activity in collaborative argumentation. In P. Dillenbourg, A. Eurelings, & K. Hakkarainen (Eds.), European perspectives on computer-supported collaborative learning (pp. 189-196) *Proceedings of the First European Conference on Computer-Supported Collaborative Learning*, Netherlands: Universiteit Maastricht.

Doble, J., & Johnson, J. (1990). *Science and the public: A report in three volumes. Volume I: Searching for common ground on issues related to science and technology.* New York, NY: Public Agenda Foundation.

Doble, J., Richardson, A., & Danks, A. (1990). *Science and the public: A report in three volumes. Volume III: Global warming caused by the greenhouse effect.* New York: Public Agenda Foundation.

Kuhn, D. (1991). *The skills of argument.* New York: Cambridge University Press.

Kuhn, D. (1992). Thinking as argument. *Harvard Educational Review, 62*(2), 155-178.

Morgan, G., & Smuts, T. (1994). *Global warming and climate change.* Pittsburgh, PA: Department of Engineering & Public Policy, Carnegie-Mellon University.

Ranney, M., & Schank, P. (1995). Protocol modeling, bifurcation/bootstrapping, and Convince Me: Computer-based methods for studying beliefs and their revision. *Behavior Research Methods, Instruments and Computers, 27*, 239-243.

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 models of social reasoning and behavior* (pp. 245-274). Mahwah, NJ: Lawrence Erlbaum.

Ranney, M., Schank, P., & Diehl, C. (1995). Competence versus performance in critical reasoning: Reducing the gap by using *Convince Me. Psychology Teaching Review, 4(*2), 153-166.

Ranney, M., Schank, P., Hoadley, C., & Neff, J. (1996). "I know one when I see one": How (much) do hypotheses differ from evidence? In R. Fidel, B. H. Kwasnik, C. Beghtol, & P. J. Smith (Eds.) *Advances in classification research: Vol. 5.* (ASIS Monograph Series; pp. 141-158, etc.) Medford, NJ: Learned Information.

Ranney, M., & Thagard, P. (1988). Explanatory coherence and belief revision in naive physics. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 426-432).

Roschelle, J. (1996). Designing for cognitive communication: Epistemic fidelity or mediating collaborating inquiry?. In D.L. Day & D.K. Kovacs (Eds.). *Computers, communication & mental models* (pp. 13-25). London: Taylor and Francis.

Scardamalia, M., & Bereiter, C. (1991). Higher levels of agency for children in knowledge building: A challenge for the design of new knowledge media. *The Journal of the Learning Sciences, 1*(1), 37-68.

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, Berkeley.

Schank, P., Hoadley, C., Dougery, K., Neff, J., & Ranney, M. (1993). *The ECHO educational program (EEP) coherent reasoning curriculum for Convince Me.* Berkeley, CA: University of California, Berkeley, Graduate School of Education.

Schank, P., & Ranney, M. (1991). The psychological fidelity of ECHO: Modeling an experimental study of explanatory coherence. *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (pp. 892-897),). Hillsdale, NJ: Lawrence Erlbaum..

Schank, P., & Ranney, M. (1992). Assessing explanatory coherence: A new method for integrating verbal data with models of online belief revision. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 599-604),). Hillsdale, NJ: Lawrence Erlbaum.

Schank, P., & Ranney, M. (1993). Can reasoning be taught? *Educator, 7*(1), 16-21.

Schank, P., Ranney, M., & Hoadley, C. (1999). *Convince Me* [revised computer program and manual]. In J. R. Jungck, V. Vaughan, J. N. Calley, S. J. Everse, P. Soderberg, E. D. Stanley, & J. Stewart (Eds.), *The BioQUEST library: Volume V, academic year 1998-1999* . 1999. San Diego, CA: Academic Press.

Schank, P., Ranney, M., Hoadley, C., Diehl, C., & Neff, J. (1994). A reasoner's workbench for improving scientific thinking: Assessing *Convince Me*. In G. H. Marks (Ed.), *Proceedings of the International Symposium of Mathematics/Science Education and Technology*. Charlottesville, VA: Association for the Advancement of Computing in Education.

Siegel, M. (1999). *Teaching science for public understanding: Developing decision-making abilities.* Unpublished doctoral dissertation, University of California, Berkeley.

Siegel, M., & Ranney, M. (in press). Developing the changes in attitude about the relevance of science (CARS) questionnaire and assessing two high school science classes. *Journal of Research in Science Teaching.*

Suthers, D., Connelly, J., Lesgold, A., Paolucci, M., Toth, E. E., Toth, J. & Weiner, A. (2001). Representational and advisory guidance for students learning scientific inquiry. In K. D. Forbus & P. J. Feltovich (Eds.), *Smart machines in education: The coming revolution in educational technology.* (pp. 7-35). Cambridge, MA: The MIT Press.

Suthers, D., & Weiner, A. (1995). *Groupware for developing critical discussion skills. In Proceedings of CSCL '95: The International Conference for Computer Support for Cooperative Learning* [On-line]. Available: http://www-cscl95.indiana.edu/cscl95/toc.html.

Suthers, D., Weiner, A., Connelly, J., & Paolucci, M. (1995). *Belvedere: Engaging students in critical discussion of science and public policy issues.* Paper presented at the AI-Ed 95, The 7th World Conference on Artificial Intelligence in Education.

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

Toulmin, S. (1958). *The uses of argument.* Cambridge, England: Cambridge University Press.

Voss, J., Lawrence, J., & Engle, R. (1991). From representation to decision: An analysis of problem solving in international relations. In R. Sternberg & P. Frensch (Eds.), *Complex problem solving* (pp. 119 - 158). Hillsdale, NJ: Lawrence Erlbaum.

Weidner, J., Ranney, M., & Steinbach, A. (1998). Using *Convince Me* to assess medical reasoning skills (and vice versa). *Proceedings of the International Conference of the Learning Sciences* (pp. 284-290). Charlottesville, VA: Association for the Advancement of Computing in Education.

Yankelovich, D. (1991). *Coming to public judgment: Making democracy work in a complex world.* Syracuse, NY: Syracuse University Press.

Yankelovich, Skelly, & White Inc. (1981). *The Mushiness Index: A refinement in public policy polling techniques.* New York: Yankelovich, Skelly, and White, Inc.

Author's Note

This article is based in part on a presentation given at the 2002 Annual Meeting of the American Educational Research Association, in Session 50.34: Thinking, Problem Solving, and Argumentation in Technology Environments. The research was supported in part by the United States National Science Foundation (Grant No. DGE-9554564).

Acknowledgments

I would especially like to acknowledge Michael Ranney for his detailed comments and overall support with this effort. Andrea diSessa provided helpful feedback for the overall research that included this study, as did Mark Christensen. I would also like to thank, for feedback and discussions, Christine Diehl, Patti Schank, Marcelle Siegel, Laura Germine, Anna Thanukos, Janek Nelson, Jennifer Garcia de Osuna, Ed Munnich, Jeanne Weidner, Todd Shimoda, Sarah Brem, and the other members of the Reasoning Group of the University of California, Berkeley. In addition, Ali Rezaei, Teresa Chen, and an anonymous reviewer provided helpful comments on the manuscript.