

Computational Tools for Modeling and Aiding Reasoning:
Assessing and Applying the Theory of Explanatory Coherence

By

Patricia Kathleen Schank

B.S. (University of Nebraska, Lincoln) 1987
M.S. (University of California, Berkeley) 1989
M.A. (University of California, Berkeley) 1991

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Education

in the

GRADUATE DIVISION

of the

UNIVERSITY of CALIFORNIA at BERKELEY

Committee in Charge:

Professor Michael A. Ranney, Chair

Professor Marcia Linn

Professor Peter Pirolli

Michael Clancy

1995

Computational Tools for Modeling and Aiding Reasoning:
Assessing and Applying the Theory of Explanatory Coherence

Copyright © 1995

by

Patricia Kathleen Schank

ABSTRACT

Computational Tools for Modeling and Aiding Reasoning:
Assessing and Applying the Theory of Explanatory Coherence

By

Patricia Kathleen Schank

Doctor of Philosophy in Education

University of California at Berkeley

Professor Michael A. Ranney, Chair

Many researchers have illustrated the difficulties and needs that children and adults have with formal and informal reasoning (e.g., Kuhn, 1989, 1993; Perkins, Allen, & Hafner, 1983). The Theory of Explanatory Coherence (TEC) and its associated connectionist model, ECHO, offers an account of how people decide the plausibility of beliefs asserted in an explanation or argument (e.g., Ranney & Thagard, 1988; Thagard, 1989). We have found that ECHO usefully predicts how people evaluate hypotheses, evidence, and other propositions regarding various situations (Schank & Ranney, 1991 & 1992). Insights from these descriptive studies led us to develop a prescriptive ECHO-based "reasoner's workbench" computer program—*Convince Me*—and an associated scientific reasoning curriculum to help students structure, restructure, and assess their knowledge about often controversial situations (e.g., Schank & Ranney, 1993). Using *Convince Me*, students can (a) easily create, modify, load, and save arguments, (b) rate how strongly they believe each statement in the argument, and (c) run an ECHO simulation to see which statements their argument helped to support or reject (and which ones it left neutral)

from ECHO's point of view. Disparities between students' own evaluations and ECHO's can help them pinpoint inconsistencies in their arguments. As a result, they may re-evaluate their beliefs, reformulate their arguments, or even adjust ECHO's numerical parameters to better model *their* way of thinking.

Does *Convince Me* significantly help students articulate and revise their theories? This dissertation discusses my (and others') prior work on modeling and aiding reasoning, then assesses the effectiveness of the *Convince Me* system. Also addressed are questions regarding the evidence/hypothesis distinction and effects of context on reasoning. Results suggest that although the distinguishing characteristics of data and theory are vague—even for experts who study scientific reasoning professionally—*Convince Me* lends a sophistication to novices' discriminative criteria across contexts, making their epistemic categorizations more expert-like both during, and after, its use. Further, more accurate and honest portrayals of these constructs as fuzzy and dependent on context may help students view science as a dynamic field that requires the continuous examination and revision of ideas, rather than the memorization of disconnected "facts."

TABLE OF CONTENTS

ABSTRACT	1
LIST OF TABLES	viii
LIST OF FIGURES	xi
ACKNOWLEDGEMENTS	xiii
1. INTRODUCTION.....	1
The Theory of Explanatory Coherence (TEC)	5
ECHO: A Connectionist Implementation of TEC	6
Related Descriptive Work	8
Philosophy	9
Plausibility and Entrenchment.....	9
Hypotheses and Evidence.....	11
Normative Models of Reasoning and Formal Rationality	12
Psychology	14
Cognitive Psychology.....	14
Social Psychology	15
Context and Reasoning.....	16
Computational Models of Reasoning	18
Probability Networks.....	18
Other Computational Models	21
Related Prescriptive Work.....	24
2. THE GROUNDWORK: DESCRIPTIVE MODELING OF EXPLANATORY EVALUATIONS WITH ECHO	30
Post-hoc Modeling.....	31
Predictive Modeling	32
Modeling Textually Embedded Propositions	32

Modeling Verbal Protocols.....	34
Extended Dynamic Modeling of Protocols and Competing Beliefs.....	37
Modeling Attention and Memory Constraints with WanderECHO ...	40
Automating Knowledge Elicitation and Supporting Argument Development with <i>Convince Me</i>	41
3. <i>CONVINCE ME</i>	43
An Example Argument.....	43
Associated Materials	49
Pre-Test.....	50
Curriculum Units	52
Unit 1, "Evidence, Hypotheses, and Theories"	52
Unit 2, "Reasoning About Arguments"	53
Unit 3, "Using <i>Convince Me</i> "	53
Integrative Exercises	54
Post-Test	54
Exit Questionnaire	55
4. PRESCRIPTIVE STUDIES USING <i>CONVINCE ME</i>	56
Study 1: Experts vs. Novices, and The Hypothesis/Evidence Distinction	56
Method.....	58
Participants	58
Design and Procedure.....	59
Results	60
Propositional Ratings.....	60
Relation Between Novices' Epistemic Categorizations and "Checkbox" Descriptions	66
Experts' Ratings of Novices' Definitions	67

Exit-Questionnaires and Comments	69
Discussion.....	70
Study 2: Determining the Efficacy of the <i>Convince Me</i> Environment.....	74
Method.....	75
Participants	75
Design and Procedure.....	75
Results	77
No Differences Between Written and <i>Convince Me</i>	
Samples.....	77
Propositional Ratings.....	78
Regression Analyses.....	81
Two-Dimensional Evidence- and Hypothesis-likeness	
Plot.....	84
Relation Between Epistemic Categorizations and	
"Checkbox" Descriptions	86
Experts' Ratings of Novices' Definitions by Group.....	87
Argument Revisions	88
Believability-Activation Correlations	90
Argument Analyses	92
Exit-Questionnaires and Comments	99
Discussion.....	102
5. SUMMARY, DISCUSSION, AND CONCLUSIONS	104
Summary.....	104
Discussion.....	105
Can Reasoning be Modeled?	105
Can Reasoning be Improved?.....	106
The Role of Context	108

The Role of Technology	109
TEC-Based Modeling and Instruction	110
Future Directions	111
Adding More Representations (e.g., Diagrams) to <i>Convince Me</i>	111
Modeling Human Processing Limitations, and Other Model Modifications	114
Collaborative Work	114
Social Versus "Scientific" Controversies, Typicality Studies	115
Conclusions	115
REFERENCES	117
APPENDIX A: Pre-Test.....	138
APPENDIX B: Unit 1, "Evidence, Hypotheses, and Theories"	158
APPENDIX C: Unit 2, "Reasoning About Arguments".....	168
APPENDIX D: Unit 3, "Using <i>Convince Me</i> "	184
APPENDIX E: Integrative Exercises (Versions Used with <i>Convince Me</i> and Written Groups).....	212
Exercises.....	213
Passages	213
Exercise 1	213
Exercise 2	214
Exercise 3	214
Exercise 4	214
Instructions	215
<i>Convince Me</i> Group.....	215
Written Group.....	216

APPENDIX F: Post-test (Questions That Differ From the Pre-test Only).....	219
APPENDIX G: Exit Questionnaire	223
APPENDIX H: New Unit 3, "Using <i>Convince Me</i> ," Extensively Revised for the New Argument Diagram/Listing Version of the Software.....	227

LIST OF TABLES

Table 1.1. Correspondence between ECHO networks and Pearl's (1988) probability networks (summarized from Thagard, in press).	20
Table 1.2. Continuum of appropriate approaches for different kinds of problems (from Thagard, in press).	21
Table 3.1. Rating instructions and examples of isolated propositions.	51
Table 3.2. Propositions embedded within a story context.	52
Table 4.1. Within-group correlations between believability and hypothesis-likeness (B-H), evidence-likeness and hypothesis-likeness (E-H), and believability and evidence-likeness (B-E), Study 1 (from Ranney et al., 1994).	61
Table 4.2. Between-group correlations regarding believability (B-B), evidence-likeness (E-E), and hypothesis-likeness (H-H), Study 1 (from Ranney et al., 1994).	61
Table 4.3. Frequency and correlational data regarding novices' checked descriptions of a statement, and their categorization of the statement as hypothesis or evidence, Study 1.	67
Table 4.4. Novices' mean pre-test definition scores, post-test change, and intercoder reliability correlations among five (expert) coders, Study 1.	68
Table 4.5. Novices' comments about the <i>Convince Me</i> system.	69

Table 4.6. Some factors that appear to influence a proposition's classification as a hypothesis or piece of evidence (adapted from Ranney et al., 1994).....	72
Table 4.7. Descriptive statistics for Written and <i>Convince Me</i> groups, all measures. Differences between groups were not significant.....	78
Table 4.8. Within-group correlations between believability and hypothesis-likeness (B-H), evidence-likeness and hypothesis-likeness (E-H), and believability and evidence-likeness (B-E), Study 2.	80
Table 4.9. Between-group correlations regarding believability (B-B), evidence-likeness (E-E), and hypothesis-likeness (H-H), Study 2.	80
Table 4.10. Full Model Regression ANOVA for believability ratings based on the predictors hypothesis-likeness, evidence-likeness, and their interaction.	82
Table 4.11. Post-hoc tests of regression slopes for each potential believability rating predictor—hypothesis-likeness (H), evidence-likeness (E), their interaction (H*E), and a constant.	83
Table 4.12. Frequency and correlational data regarding novices' checked descriptions of a statement, and their categorizations of a statement as hypothesis or evidence, Study 2.	86
Table 4.13. Novices' mean pre-test definition scores, post-test change, and intercoder reliability correlations among four (expert) coders, Study 2.	88
Table 4.14. Changes to arguments and ratings, <i>Convince Me</i> and Written groups.	89

Table 4.15. Overall belief-activation correlations on the first argument, the last revised argument, and all arguments.....	91
Table 4.16. Believability-activation correlations and the mean number of hypotheses, evidence, new propositions (hypotheses or evidence that were not in the given text/situation), explanations (including joint explanations, in parentheses), and contradictions for novices' arguments—overall (across tests and integrative exercises), and for the tests and integrative exercises separately.....	93
Table 4.17. Believability-activation correlations for all novices, overall without the abortion case, and on the "abortion" argument (for their last revision), believability ratings for the "abortion is okay" proposition versus the "abortion is wrong" proposition, number of pro ("okay") versus con ("wrong") abortion propositions, and total number of argument propositions and links.....	96
Table 4.18. Number of "my side" versus "other side" propositions for the abortion argument, by group. (Differences are not significant, $X^2 = .70$.)	97
Table 4.19. Questionability of evidence/hypothesis categorization for the "abortion" argument.	98
Table 4.20. How much novices thought they learned (on a 1-to-7 scale, in which 1 = not much and 7 = a lot).....	100
Table 4.21. <i>Convince Me</i> users' comments about the system, after using it.	100
Table 4.22. Written students' comments about the system, after reading about it. ...	101

LIST OF FIGURES

Figure 2.1. Topology of two conflicting explanatory theories {H0, H1, etc. vs. H2, H3, etc.}.....	33
Figure 2.2. The "bifurcation/bootstrapping method" (from Schank & Ranney, 1992).....	35
Figure 2.3. Mean pendular-release path believability ratings for the various alternatives, on a 1 to 9 scale, prior to feedback (from Schank & Ranney, 1992).....	36
Figure 3.1. Two viewpoints about freezing ice cubes (sample topology).....	44
Figure 3.2. A user adds and classifies a belief about the speeds at which water of different initial temperatures freeze (bottom) in response to <i>Convince Me's</i> feedback (middle). (Cf. the diagrammatic interface in Chapter 5, "Future Directions.")	46
Figure 3.3. Adding an explanation.	47
Figure 3.4. Adding a contradiction.....	47
Figure 3.5. Rating a statement's believability.....	48
Figure 3.6. Modifying ECHO's parameters (default values are shown).....	48
Figure 3.7. Associated materials and the intended sequence of use.....	50
Figure 4.1. Summary of this experiment's method.....	59

Figure 4.2. Novices' and experts' correlation distributions, Study 1 (from Ranney et al., 1994). H-E, H-B, and B-E refer to hypothesis-evidence, hypothesis-believability, and believability-evidence correlations, respectively.....	63
Figure 4.3. Summary of this experiment's method.....	76
Figure 4.4. Plot (and listing) of propositions' mean ratings along the hypothesis-likeness and evidence-likeness axes, with mean believability (in parentheses), Study 2.....	85
Figure 4.5. Argument and rating change episodes, <i>Convince Me</i> and Written groups.	90
Figure 4.6. Overall model's fit, all arguments, <i>Convince Me</i> and Written groups (from Schank & Ranney, 1995).....	92
Figure 5.1. Adding a belief to the ice cubes argument (cf. Figure 3.2) in response to <i>Convince Me's</i> feedback. This modified version of the software displays (a) an argument diagram (upper right) rather than merely displaying the "activational thermometer" icons in rows and columns, and (b) a listing of all explanations and contradictions (below the left side of the diagram).	113

ACKNOWLEDGEMENTS

This dissertation, like many, could not have been written without the support, advice, and encouragement of many people over the years. Special thanks go to my committee members, especially Michael Ranney and Marcia Linn, who mentored me during my graduate years and offered very helpful comments and suggestions on early drafts of this thesis (not to mention many other papers!).

I've been lucky to work with many excellent faculty during my graduate and undergraduate years—including Marcia Linn, Michael Clancy, Peter Pirolli, Lawrence Rowe, David Wessel, Stephen Palmer, Melvin Thornton, Susan Wiedenbeck, and Stuart Margolis. But above all, I feel especially fortunate to have had the opportunity to work with my advisor, Michael Ranney. I simply couldn't imagine a better role model, mentor, collaborator, and friend. His knowledge, thoughtfulness, and rigor always impressed me, and along with his great personality and sense of humor, helped me develop a great deal of joy and pride in my work.

I'm also very grateful for loving family and friends (Mom and Dad, Krysten, the Chrises, Maxine, Sherry, Wanda, Dave, Shirley, and many more) and a supportive EMST community. I especially want to thank Larry Hamel for his love, encouragement, massages, and many other fun distractions (singing, salsa lessons, the wonderful trip to Kauai...) that made my last year in graduate school the best!

This thesis is dedicated to my brother and sister, Jeff and Sandy, for their constant love and encouragement, and our ever-engaging discussions (often related to reasoning!) throughout my life.

1. INTRODUCTION

How do people decide what to believe and disbelieve in a situation? How do they assess the plausibility of an assertion in an argument? Do people agree on what is evidence and what is hypothesis? How does context affect reasoning? Can we improve human rationality? This work attempts to answer these questions by integratively using including cognitive modeling, experimental studies, and the development of software and instructional curricula. More specifically, a theory-based computational model was used to predict how people would evaluate hypotheses and evidence in various (often controversial) situations. Insights from this work led to the development of instructional software and curricula that focused on methods of scientific reasoning. This report describes the design, use, and assessment of these computational tools.¹

¹The following are definitions for common terms used throughout this report:

Belief: A proposition, i.e., a hypothesis or piece of evidence.

Hypothesis: A proposition that explains or predicts something of interest. One possible inference, opinion, or view; some reasonable people may disagree.

Evidence: A proposition that seems based on "objective-like" criteria; for example, an acknowledged common fact or statistic, or a reliable memory or observation.

Explanation: A relation among beliefs that makes clearer or understandable another belief; it shows how or why something happened. The coordination of beliefs such that some are accounted for (often causally) by others.

Contradiction: The relation between a pair of beliefs that are mutually exclusive or (at least) unlikely to both be true.

Theory: A system of evidential and hypothetical beliefs that have a unifying theme.

Not surprisingly, people often differentially evaluate the plausibility of similar (or even identical) beliefs when reasoning about complex situations. They generally hold their beliefs as long as they help explain many of their experiences, even if these beliefs don't fit precisely into a scientific framework (i.e., the beliefs may conflict with established scientific hypotheses). For example, students learning physics tend to hold strong intuitive beliefs about the physical world that tend to resist revision (e.g., diSessa, 1983 & 1993; Ranney 1987/1988; Hartley, Byard, & Mallen, 1991).

Ranney and Thagard (1988) characterize belief evaluation and revision as the result of seeking explanatory coherence between theories and observations, in which the plausibility of a belief generally increases with its increasing simplicity (e.g., fewer necessary co-hypotheses), increasing breadth (i.e., more coverage of observation), and decreasing competition with alternate (especially entrenched) beliefs (cf. Johnson & Smith, 1991). These principles (along with others) play important roles in evaluations of the quality of an explanation (Schank & Ranney, 1991; Read & Marcus-Newhall, 1993), and together comprise the Theory of Explanatory Coherence (TEC; e.g., Thagard, 1989). This theory guides the research reported here.

Thagard (1989) and Ranney (in press) describe a computational implementation of TEC, called ECHO. The ECHO model is based on the claim that beliefs are related explanatory entities, and evaluating their plausibility is an interactive, principled, coherence-seeking process (Thagard, 1989; Ranney, in press).²

Argument: A system of beliefs that is generally more complex than one explanation/contradiction, but less than that of a theory.

² ECHO's "theoretical/systemic" coherence differs from (and is generally orthogonal to) standard notions of "linguistic" coherence (Ranney, Schank, & Ritter, 1992; Schank & Ranney, 1992). In ECHO, coherence is seen from the perspective of competing theories, where the dynamic tension represented as explicitly conflicting

Belief evaluation in ECHO involves the satisfaction of many constraints, determined by the explanatory relations among propositions and a few processing parameters. Although "optimal" reasoning may not be possible by machines or humans within a reasonable time (Dreyfus, 1992; Winograd & Flores, 1987; Tash, 1994; Verbeurgt & Thagard, forthcoming), ECHO provides an efficient approximation (Verbeurgt & Thagard, forthcoming; cf. Tash, 1994). More complete descriptions of ECHO's algorithms are given below, and elsewhere (e.g., Thagard, 1992).

ECHO has been used in the past to model juror reasoning (Thagard, 1989) in which explanatory coherence plays a crucial role (Pennington & Hastie, 1988), to understand mental models of social interactions and relationships (Miller & Read, 1991), and to model scientific reasoning (Thagard, 1989, 1992). Ranney and Thagard (1988), in the first application of ECHO to modeling experimental data, simulated (*ex post facto*) changes in subjects' beliefs and their conceptions of physical motion (Ranney, 1987/1988; cf. Nersessian, 1989). Early on, we hypothesized that if ECHO usefully models human reasoning, it might also prove useful as a tool for teaching coherent argumentation (e.g., Ranney, in press; Ranney, Schank, Ritter, & Carlock, 1991). We found ECHO helpful for predicting and interpreting subjects' reasoning patterns (Schank & Ranney, 1991 & 1992), and assessing the descriptive aspects of ECHO gave us insight into its prescriptive utility (cf. Ranney, in press). Encouraged

theories reduces the overall coherence of a system of propositions (compared to a network dominated by a single-theory). In contrast, textual/discourse coherence is generally viewed as increasing with more explicit relations among various entities and assertions in a text, and less reliance on implicit background knowledge for making inferences (such as anaphoras; e.g., Givon, 1991; Trabasso, van den Broek, & Suh, 1989). (E.g., the textual stimuli used in Schank & Ranney, 1991, were designed to be low in systemic coherence and high in linguistic coherence.)

by these results, insights, and the engagement of Margaret Carlock's (1990) adolescent students using her "interactive front end" (IFE)³ for ECHO, we developed *Convince Me*, a computer-based "reasoner's workbench" based on ECHO (implemented in HyperCard; Schank & Ranney, 1993). We also designed an associated ECHO-based reasoning curriculum that addresses documented weaknesses in reasoning, using several scientific and everyday controversies (e.g., regarding competing theories of motion, continental drift, animal behavior, freezing materials, abortion, the legalization of drugs, etc.). While other formal systems exist for the analysis and generation of arguments (e.g., VanLehn, 1985), there seems to be no other that is based upon a particular processing model, or that includes a computational model that actually yields predictions about the plausibility of an argument's assertions (e.g., for the benefit of students).

This study assesses the utility of employing *Convince Me* to generate and analyze scientific arguments. Does using the system help students gain a better understanding of hypotheses, evidence, and theories, and construct more coherent arguments?

The remainder of Chapter 1 describes TEC and ECHO in more detail, and reviews related work. Chapter 2 summarizes descriptive studies involving ECHO. *Convince Me* and its associated curriculum are described in Chapter 3. Chapter 4

³ Using Carlock's IFE, students could categorize given statements as hypotheses or evidence, indicate which statements explain and/or contradict each other, and run ECHO simulations of arguments. Among other things, *Convince Me* extends the IFE by allowing students to enter their own beliefs, querying students for their evaluations, providing "model's fit" feedback, offering enhanced graphical input and output tools, and providing on-line help.

summarizes two studies involving *Convince Me*. Finally, Chapter 5 offers a summative discussion and proposes future work.

The Theory of Explanatory Coherence (TEC)

In TEC, coherence involves relations among two or more propositions (beliefs) that may “hold together” or “resist holding together.” The following principles establish the local pairwise relations among cohering and incohering propositions (selectively from Schank & Ranney, 1991, Thagard 1989 & 1992, and Ranney & Thagard, 1988):

- (1) *Symmetry*: Coherence and incoherence are symmetric relations between pairs of propositions.
- (2) *Explanation*: A proposition that independently explains another proposition also coheres with it. Propositions that together explain a proposition cohere with each other and with the explained proposition.
- (3) *Simplicity*: The plausibility of a proposition is inversely related to the number of co-propositions it needs to help explain a proposition.
- (4) *Data Priority*: Results of observations, such as evidence and acknowledged facts, have a bias toward acceptability.
- (5) *Contradiction*: Contradictory propositions incohere.
- (6) *Competition*: Two propositions may incohere/compete if they explain a third proposition (an explanandum), yet are not themselves explanatorily related (cf. Harman, 1989)—suggesting that the two propositions are probably mutually exclusive. (This principle is optionally invoked in a variant of ECHO, called ECHO2; see Ranney, Schank, Mosmann, & Montoya, 1993, Thagard, 1991a, and below.)

(7) *Acceptability*: The acceptability of a proposition depends on its coherence with the system of propositions in which it is embedded. A proposition's acceptability increases as it coheres more with other acceptable propositions and *incoheres* more with *unacceptable* propositions. In ECHO, a proposition's acceptability is measured by its activation value, ranging from -1 (complete rejection) to 1 (complete acceptance).

Schank and Ranney (1991) and Read and Marcus-Newhall (1991) show that these principles (save competition) play important roles in explanations. They found that subjects prefer explanations that account for more data, are simpler, and involve beliefs that can be further explained. Subjects' evaluations of explanations are also changed by the availability of competing explanations. Ranney et al., (1993), however, assessed the predictive utility of the competition principle (see Chapter 2), and suggest that it needs refinement.

ECHO: A Connectionist Implementation of TEC

ECHO uses a connectionist architecture in which each node represents a proposition. Hypothesis evaluation is treated as the satisfaction of multiple constraints derived from the explanatory relations, and several parameters provide degrees of freedom (Thagard, 1989, 1992). By itself, ECHO neither "learns" connection weights nor infers new propositional relationships; these are provided, depending upon the methodology employed, by default, by the experimenter, or by the subject (see Chapters 2-4). The number of parameters used in computer models often leads to questions like: How arbitrary are the simulation findings? How stable is the model under a given set of parameters (i.e., does one set work over a wide range of tasks)? Should some parameters be regarded as variant or invariant over tasks? Thousands of simulations have been used to determine the impact of parameter

settings on ECHO's performance, and the values of four parameters—the *excitation*, *inhibition*, *decay*, and *data excitation* rates—appear most critical (Thagard, 1989; Ranney & Thagard, 1988). However, the criteria used to decide the default values of these parameters were initially practical rather than empirical. Ranney and Thagard (1988) and Schank & Ranney (1991) provide the first empirical backing for assigned parameter values.

The excitation value determines the weights on links between cohering propositions, thus implementing the explanation principle (2). The inhibition value determines the weights on links between incohering propositions, implementing the contraction and competition principles (5 and 6). The data excitation value specifies the weight on links between data (usually evidence) and the “special evidence unit” (SEU, a unit with activation set at a constant 1.0) to implement the data priority principle (4). The decay value specifies the percentage of the (absolute) activation that a proposition loses at each cycle. Thagard (1989) describes interesting psychological interpretations of these parameters: The *tolerance* of the system is the absolute value of the ratio of excitation to inhibition; highly tolerant systems may “believe” several competing hypotheses, while systems with low tolerance tend to accept a single theory. Decay is the *skepticism* of the system; as decay increases, asymptotic activation values of propositions will be compressed toward zero/indeterminacy (although the simulations will not be strikingly qualitatively different). Skepticism applies directly to all propositions, while tolerance is most obviously relevant to contradictory or competing propositions.

Given declared input propositions and relations among them, node activations are updated using a simple settling scheme. ECHO creates units to represent the propositions, and sets up the following links (with additive link weights): (1) excitatory links between propositions (including co-hypotheses) that are explanatorily or analogously related; (2) inhibitory links among contradictory propositions; (3)

inhibitory links between (nonevidential) competing propositions (in ECHO2 only; implemented by searching the network representation for propositions that independently explain a third proposition, which are then marked as competing); (4) excitatory links between evidence and the SEU (with the data excitation weight). (The data priority of particular evidence may be specified separately, if desired.) If there are unexplained data, the decay rate is increased appropriately (i.e., unexplained evidence raises skepticism). Unit activations are updated in cycles until the network settles and the change in all units is asymptotic. At each cycle, the activation of a particular unit u_j is determined by the decay rate and the net input to the unit (the weighted sum of the activation of each neighboring unit u_i , where a weight is the common link value, w_{ij}), and is updated using the following equation (cf. Rumelhart & McClelland, 1986):

$$u_j(t+1) = u_j(t) (1 - \text{decay}) + \left\{ \begin{array}{ll} \text{net}_j (\text{max} - u_j(t)) & \text{if } \text{net}_j > 0 \\ \text{net}_j (u_j(t) - \text{min}) & \text{otherwise} \end{array} \right\}$$

$$\text{where } \text{net}_j = \sum_i w_{ij} u_i(t)$$

Related Descriptive Work

This section focuses on how TEC complements and differs from existing models of reasoning. For instance, how is the TEC related to other theoretical and computational models of reasoning? Does the ECHO model help us understand reasoning in ways the others do not?

Not surprisingly, a variety of models of cognition have been proposed, many computational (rule-based, connectionist, and hybrid; e.g., Anderson, 1987; Johnson-Laird, 1988; Kintsch, 1988; Laird, Rosenbloom, & Newell, 1986; Holland, Holyoak, Nisbett, & Thagard, 1989; Thagard, 1989), and many non- or less-computational

(e.g., Lakoff & Johnson, 1980; Kolers & Smythe, 1984). This plethora of models has also led some researchers to ask more meta-level questions, such as: What kinds of questions do the underlying metaphors and methodological tools inspire, and implicitly restrict? How have mental metaphors changed over time? How much of this change is explained by new technology? Several pieces discuss the emergence, implications, and limitations of a range of metaphors for the mind (e.g., Gentner & Grudin, 1985; Gigerenzer, 1991; Lakoff & Johnson, 1980; McCloskey, 1991; Searle, 1990), and although interesting, these issues are beyond the scope of this work.

Several attempts to specifically account for the inter-relationships, revisions, and structure of individuals' beliefs have been advanced—including schemata and mental models, conceptual maps, discriminability analyses, and probabilistic belief networks (cf. Austin & Shore, 1993; Bartlett, 1932; Carey, 1985; Chi, Feltovich, & Glaser, 1981; Gentner & Stevens, 1983; Pearl, 1988). These accounts have significant methodological or pragmatic limitations. For instance, "concept maps," popular in contrasting (relative) experts with novices, are commonly post-hoc, seat-of-the-pants analyses by theorists who are "eyeballing" their data; even researchers who attempt to explicitly contrast the knowledge structures of two individuals or groups rarely, if ever, report inter-coder reliability measures (cf. Ranney, in press). At another extreme, Bayesian-style probability networks have more rigor, but reasonably-sized networks require many estimates of (e.g., conditional) probabilities that humans cannot or have not pondered (see below, and Thagard, in press).

Philosophy

Plausibility and Entrenchment

Over 40 years ago, the philosopher Nelson Goodman (1954/1983), argued that the question, "How does one distinguish generalizations that are warranted from those that are not?," resists formal solution. He reduced several problems, including the

problem of inductive inference, to the problem of confirmation: When does a set of evidence confirm a particular hypothesis? Goodman introduces this "new riddle of induction," the problem of distinguishing "law-like" from accidental (non-law-like) hypotheses, with his famous "grue and green" example⁴. In a classic attempt to 'solve' the problem, Goodman posed a theory of *projection* for ordering hypotheses and selecting among inferences. Goodman's theory included a principle of "entrenchment" (among others), where the entrenchment of a proposition is viewed (roughly) its validity or utility, determined by the proposition's record of past projections. His (circular) theory states that a predicate acquires entrenchment from other predicates related to it; i.e., hypotheses that are made up of more entrenched predicates are themselves better entrenched, and hence more valid. Thus, Goodman reduces the problem of distinguishing warranted from unwarranted generalizations to that of identifying the hypotheses that are "more entrenched." Similarly, Quine and Ullian (1970) list breadth and compatibility with previous beliefs (along with simplicity and refutability) as the main virtues that count toward the plausibility of a hypothesis.

TEC provides a holistic theory to account for how warranted and unwarranted explanatory hypotheses can be distinguished from one another. The theory also

⁴The grue and green example is posed thus: consider the predicate grue, which applies to any blue thing, and also to anything which was examined before some given time t and found to be green. Suppose all emeralds examined before time t are green. Then both hypotheses, (1) all emeralds are grue, and (2) all emeralds are green, are both equally confirmed by the evidence (Goodman, 1954/1983). These hypotheses compete with each other if we want to accept one of them as the *best* explanation of the data (e.g., Harman, 1989).

entails a concrete, computational model intended to capture part of how people reason, making some conceptual connections between philosophy and artificial intelligence more clear (Thagard, 1989, 1991b, 1992). In ECHO, other things being equal, the global coherence of a set of propositions increases as the number of propositions and explanatory relations between them increase. In Goodman's terms, one might say that a proposition becomes more valid as the proposition becomes entrenched. The entrenchment of a proposition can be described as depending on both the activation of the proposition and the coherence of the system of propositions within which the proposition is embedded. In other words, as the coherence of a network increases, the propositions in the network that have the highest activations are generally the most entrenched. In Quine and Ullian's terms, this roughly corresponds to the propositions that have the most breadth, simplicity, and compatibility with other beliefs.

Hypotheses and Evidence

TEC is also relevant to the philosophy of science, law, epistemology, and metaphysics (see Thagard, 1989). For instance, in scientific reasoning, the distinction between evidence and hypothesis (or theory) appears to be fundamental, a distinction that Kuhn (1989) claims "separates children from scientists" (but see other views below), and many textbooks imply that the distinction is clear and apparent to experts (e.g., Giere, 1991). Our own simulations with ECHO (e.g., Ranney et al., 1993; Schank & Ranney, 1991) also rely on the evidence/hypothesis contrast (as do other studies that have assessed principles of explanatory coherence, cf. Read & Marcus-Newhall, 1993). TEC includes the principle of data priority, which essentially holds that a piece of evidence is more acceptable than a mere hypothesis, *all other things being equal*. ECHO reifies this principle by linking connectionist nodes that represent evidence directly to the model's source of activation; in contrast, nodes

representing hypotheses in ECHO are only indirectly linked to the activational source.

Philosophers of science (e.g., Feyerabend, 1978; Hanson, 1958/1965) and others have argued, though, the data/theory classification may not be clear-cut—that is, observations and experiments are often “loaded with the theories” (Hanson, 1958/1965, p. 157). In a similar vein, Popper (1978) answered the chicken-and-egg question of “Which came first, evidence or hypothesis?” with: “an earlier kind of hypothesis”—by which he explains that humans never engage in theory-free observations, although their theories may of course undergo revision as a result of new evidence (also cf. Tweney, Doherty, & Mynatt, 1981). Feminist critiques (e.g., Longino, 1990; Keller, 1985) also argue that observation presupposes interpretation—that it is a “near truism that there is no such thing as raw data” (Keller, 1985, p. 130). Study 1, in Chapter 3, further (and empirically) explores the evidence/hypothesis distinction.

Normative Models of Reasoning and Formal Rationality

There has been a long history of attempts to use normative theories to model (and prescribe) rational judgment and decision making. These theories include formal logic (deduction, conditional reasoning, etc.), hypothesis evaluation via Bayes's theorem, and expected utility and game theory, among others (e.g., Boole, 1854; Osherson, 1990; Slovic, 1990). Until the 20th century, reasoning was all but equated with logic. Since then, however, many have argued that formal methods yield a misleading view of reasoning (e.g., Toulmin, 1958), and others have demonstrated discrepancies between human performance and the prescriptions of logic. These include atmosphere effects in reasoning about quantifiers (Woodsworth & Sells, 1935), confirmation bias and difficulties with *modus tollens* (e.g., Wason & Johnson-Laird, 1972; Wason, 1977; Stich, 1990), and biases in probability estimates

and base rates (e.g., Tversky & Kahneman, 1976; Osherson, 1990). Despite these discrepancies, some still believe that humans reason by some sort of logic (e.g., Rips, 1983), while others suggest that we use heuristics (e.g., availability, representitiveness; Tversky & Kahneman, 1976), mental models (essentially, evaluating conclusions by reference to concrete examples that satisfy given premises; Johnson-Laird, 1983), or pragmatic inference schemata (e.g., using abstract rules about permission; Cheng & Holyoak, 1985).

Formal methods are problematic in that they usually require an enormity of calculations by agents with limited computational resources (i.e., people, and also computers). Simon (1955) introduced the concept of bounded rationality, claiming that unlike normative models that optimize (e.g., Bayesian decision theory, probability networks, etc.), humans "satisfice." Whereas ECHO does not explicitly model human computational power and memory limitations (e.g., it will run on a network of almost unlimited size and will continue updating all activations until the entire network settles; Ranney, in press), Thagard (in press) argues that it is more practical, tractable, and better at modeling human performance than its probability-based relatives (see "Probability networks" below). WanderECHO (Hoadley, Ranney, & Schank, 1994) was developed to explicitly model aspects of such constraints (see Chapter 2).

Tash (1994) and others (e.g., Dreyfus, 1992; Winograd & Flores, 1987) claim that "optimal" reasoning cannot be automated due to computational and other limitations (e.g., the embedded, situated nature of being-in-the-world implies the need for "infinite context"). Similarly, Thagard (forthcoming) argues that no computer, including the human brain, could perfectly compute coherence within a reasonable time. However, Tash (1994) and Thagard (forthcoming)—unlike the others—don't reach the "negative" conclusion that therefore reasoning cannot be automated. They agree that it can't be *optimal*, but people don't reason optimally (e.g., have infinite

context), so why should a program? Efficient approximations can work effectively in most cases, and do work fairly well for humans. Russell and Wefald (1991) reformulate Simon's bounded rationality concept to one of "limited rationality." First, they argue that the rationality of an agent is best seen as *relative* to another agent—the "designer"—that judges or directs it, relieving the agent of the burden of decision-theoretic computation. The designer can then choose the agent among a set of agents that "performs best" (akin to "natural selection" or genetic algorithms; Holland, 1992). Although this appears to simply shift the problem from the agent to the designer, Tash (1994) argues that the best way to automate reasoning is to use such a metalevel architecture. Tash and Russell (1994) offer an example of such a system. Analogously, one could apply a meta-level architecture to ECHO, creating a designer-agent that encodes a situation into alternate ECHO networks, and chooses the "best" set of evaluations among these networks based on some criteria (or one could have multiple human encoders encode a single protocol, as in Schank & Ranney, 1992, and contrast and select the "best" from the resulting simulations).

Psychology

Cognitive Psychology

Cognitive psychologists such as Anderson and Bradshaw (Anderson, 1983, 1985; Bradshaw & Anderson, 1982) have dealt primarily with how information is *recalled* from memory and not on how hypotheses and evidence are *evaluated* per se, but their underlying models are similar to ECHO on one level. For instance, Anderson (1983, 1985) suggests that as one is forced to elaborate and explain information, one creates 'links' to the to-be-remembered items, creating alternate retrieval routes. To explain this phenomena, he proposes a "spreading activation" model of memory. According to this model, (a) strongly encoded information receives greater activation through "associative priming," (b) the more links that are

created to an item in memory, the more likely it is to be remembered, and (c) highly activated information is recalled into short-term memory for use. (Cf. the robust “priming effect” that follows from this model; Meyer & Schvaneveldt, 1971; McKoon & Ratcliff, 1980).

The underlying memory network proposed by Anderson is similar to an ECHO network in that they are both networks through which activation spreads—but they are quite different with regard to what a network's elements represent (e.g., memory-level vs. believability) and the types of relations conceived between elements. ECHO can be considered an extension of the spreading activation metaphor to hypothesis evaluation, particularly to unconscious evaluation (Ranney, in press).

Social Psychology

In reaction to the traditional associationist model of memory (e.g., Thorndike, 1922), social psychologists proposed higher-order structures such as goals and schemata to explain how people draw inferences and expectations about situations. Several researchers (e.g., Brewer & Treyens, 1981; Hastie, 1980; Bransford & Johnson, 1972; Fiske & Taylor, 1984) have shown that people tend to remember schemata-congruent information, forget schemata-irrelevant information, and have intrusions that are congruent with schemata. More rigorous information-processing models for social perception and memory are few, however (Hastie & Carlston, 1980; Wyer & Srull, 1989).

TEC is relevant to social psychology in that it predicts the following: "If a proposition is highly coherent with the beliefs of a person, then the person will believe the proposition with a high degree of confidence. If a proposition is incoherent with beliefs of a person, then the person will not believe the proposition" (Thagard, 1989). Further, TEC offers an alternative model (ECHO) for understanding

how people draw inferences in the social environment—one that yields precise, testable predictions. For example, people might be said to accept hypotheses about others on the basis that these hypotheses yield coherent explanations of behavior (Thagard, 1989). Pennington and Hastie (1988) claim that explanatory coherence plays a crucial role in jurors' decision making, and Read, Miller, and Marcus-Newhall (Read & Marcus-Newhall, 1993; Miller & Read, 1991) have used the principles of explanatory coherence to understand and guide models of social interaction and relationships.

Context and Reasoning

As mentioned above (see "Normative Models of Reasoning and Formal Rationality"), reasoning was all but equated with logic until the 20th century. Even Piaget's monumental work (e.g., Piaget, 1970) focused on strategies required for formal reasoning, but not on the situational context. While Piaget noted that the context of a task seemed to influence reasoning and conceptual change, he viewed these influences as unsystematic and thus did not include them in his theoretical work. Many teaching methods also implicitly assume that knowledge and reasoning can be abstracted from the instructional situation and applied; indeed, the transfer of knowledge and reasoning to new situations seems to be one of the primary objectives of education.

Other researchers have investigated and documented effects of context on reasoning. The "atmosphere effect" is one example of the influence of context (Woodsworth & Sells, 1935): When reasoning about syllogisms, subjects' evaluations of the validity of a conclusion vary by the kinds of quantifiers (or "atmosphere") used in the premises. For instance, if the quantifier "some" is used in one of the premises, they are more likely to state as valid a conclusion that contains the quantifier "some," even if the conclusion is invalid. This effect is robust, and

explains some discrepancies between formal, context-free reasoning and human performance. Many other researchers have reported context effects on reasoning (e.g., on critical thinking skills, Brown & Campione, 1990; on probability estimates and base rates, Tversky & Kahneman, 1976; on concepts in physics, Linn, Clement, & Pulos, 1983; on concepts in mathematics; Saxe, 1988). Still others argue that the situatedness of knowledge and reasoning calls for a completely new perspective on education—one that honors the situated nature of knowledge and makes deliberate use of context (e.g., via cognitive apprenticeship; Brown, Collins, & Duguid, 1989; Brown & Campione, 1990; see also Lave & Wenger, 1991). Researchers have also been criticized for using unrealistic contexts in their studies of reasoning (e.g., Linn, 1990, regarding Kuhn, Amsel, & O'Loughlin, 1988).

The ECHO model is domain-independent in that it evaluates the plausibility of propositions in an argument given only the argument's structure—the model doesn't "understand" the content of the argument. This might be viewed as a strength of the model, in that it can be applied to any task—indeed, ECHO has been used to successfully predict people's evaluations in a variety of domains (e.g., Schank & Ranney, 1991 & 1992; Miller & Read, 1991). Example situations have ranged from fictional, "decontextualized" controversies (e.g., regarding identifying an atomic particle as being a "zipton" or "blinkon," or identifying which of two patients has the fictional disease "glumpis") to more ecologically realistic, even visceral, controversies (e.g., regarding whether or not it is safe to send a child to a school where there is another HIV-positive child). We have found that subjects clearly consider extraneous background knowledge in their deliberations—for both the fictional controversies (e.g., Schank & Ranney, 1991) and especially the more visceral ones (e.g., Ritter, 1991). Although ECHO was largely successful in its modeling throughout the range of fictional-to-realistic situations, its success generally increases with the amount of information it has about an individual's knowledge base

(e.g., the model generally shows an improving fit over time as subjects explicate their reasonings regarding pendular-release trajectory predictions; see below, as well as Ranney et al., 1993). In this way, the model does rely on (and use) context and background knowledge.

Finally, ECHO (in *Convince Me*) is able to give feedback on how well one's propositional believability ratings reflects his or her argument's structure—but not on the argument's semantic content. The more accurate and complete a representation *Convince Me* has of the individual's argument (i.e., the context), the more accurate and helpful its feedback is likely to be. The results reported here (see later chapters) indicate that both context and strategies are important in reasoning, and *Convince Me* can make its users better reasoners with its structure-based feedback and knowledge-articulation features. Effects of context on students' abilities to discriminate between the notions of evidence and hypothesis were also investigated, and were generally (but not always) found to heighten the distinction.

Computational Models of Reasoning

Probability Networks

ECHO networks are often directly compared with probability networks (e.g., Thagard, 1989, in press). Cognitive psychologists and computer scientists have often examined beliefs and hypothesis evaluation from a Bayesian viewpoint (e.g., Anderson, 1983; Andersen, Jensen, Olesen, & Jensen, 1989; Pearl, 1988), but actual student modeling is a relatively new area (e.g., Sime, 1993; Villano, 1992). The main disadvantage of the Bayesian modeling approach has been that the experimenter (versus the subject) has provided the estimates of the relevant prior and conditional probabilities. In contrast, ECHO can be used for predictive, dynamic modeling of students' reasonings with little or no experimenter interference (as described in Chapter 2).

Thagard (1989) argues that the use of probabilities to understand human hypothesis evaluation begs the question: non-statistical theory evaluation abounds in everyday life (Tversky & Kahneman, 1976), and our probability judgments are determined *by* our judgments of the merits of alternate explanations, more so than the other way around. Thagard (1989) addresses some potentially unfavorable comparisons of ECHO with probability theory (e.g., by Feldman, 1989; Cohen, 1989; Papineau, 1989; same issue), as follows:

- Feldman (1989) says that, unlike logic and probability theory, TEC is missing a formal semantic foundation (i.e., for the weights and activity levels). Thagard argues that probability theory doesn't have a clean formal semantics either—should probabilities be interpreted as frequencies, propensities, or subjective degrees of belief? Thagard claims that probability's (apparent) superiority is derived more from its familiar syntax than from foundational advantages.
- Cohen (1989) claims that TEC doesn't provide a way to determine the acceptability of a conjunction. Thagard agrees that this problem is beyond the scope of TEC, but that it isn't solved by probability theory either, since calculating the probability of a conjunction requires knowing the degree of dependence of the conjuncts, which is often indeterminate.
- Papineau (1989) claims that people can learn to reason better probabilistically. Thagard replies that sure, we should exploit probabilistic reasoning whenever possible, but we should also quit pretending—probabilities are sparsely available! "...probabilism reigns supreme as the epistemology of Eternal Beings. But explanationism survives as epistemology for the rest of us." (Thagard, in press).

Thagard (in press) shows how in principle, one can translate an ECHO network into a Pearl network, since one might interpret link weights in ECHO as corresponding to conditional probabilities, and initial node activations correspond to prior probabilities (see Table 1.1, summarized from Thagard, in press). However, problems arise in practice; for example, not all conditional probabilities can be derived from the link weights, and clustering competing and co-hypotheses can be combinatorially disastrous.

Table 1.1. Correspondence between ECHO networks and Pearl's (1988) probability networks (summarized from Thagard, in press).

	<u>in ECHO</u>	<u>in Pearl's probability nets</u>
Nodes represent...	propositions	variables
Node value...	activation [1,-1]	BEL vector (0, 1)
	initially 0 (neutral)	prior probability
Edges represent...	relative coherence/incoherence	dependencies
Edge weights	explanation, contradiction, data priority parameters	conditional probabilities
Directedness...	symmetric	directed (A causes B)
Loops...	many	eliminated (to avoid combinatorial explosion)
Additional update...	none	lambda, pi

Thagard claims that ECHO is better than Pearl networks in that it (a) is more practical, (b) is tractable, unlike Pearl networks (unless they use a "no loop constraint;" also note that ECHO networks can oscillate indefinitely in theory, but this is generally controlled by parameter choices), (c) doesn't require an assumption of independence (required by Pearl networks), and (d) takes complications (like lack of availability of probabilities) in stride, unlike Pearl networks, whose advantages are "weak in practice" since probabilities are sparsely available. However, Pearl (1986) and others argue that the "neo-probabilistic" approach reduces such difficulties for

probability networks by using local computation methods, allowing conditional probabilities to be modeled by parametric techniques, and using constraint-propagation. Sounds familiar—and not unlike ECHO! The computational load is still great, though (presumably greater than that for ECHO), and it's not clear that the results are more reliable.

In addition to showing how ECHO can use probabilistic information, Thagard (in press) claims that ECHO's final activations are usually qualitatively similar to probabilistic evaluations (although he only offers one example to support this claim). He further argues that whether one chooses to use ECHO or probability networks depends on one's priorities: For theoretical reliability, probability nets may be the better choice; but for power, speed, and human-like performance, ECHO is the clear winner. He also offers a continuum of appropriate approaches for different kinds of problems (see Table 1.2). In sum, Thagard advises researchers to not "muddy the clear probabilistic waters" if probabilities are known (e.g., from experimental studies)—that is, they should then use Pearl nets. But if probabilities are not available (as it is with the most “interesting real” cases), try ECHO.

Table 1.2. Continuum of appropriate approaches for different kinds of problems (from Thagard, in press).

EXPLANATIONIST	-----Most appropriate approach-----				PROBABILISTIC
social reasoning	scientific reasoning	legal reasoning	medical diagnosis	fault diagnosis	games of chance

Other Computational Models

There are many other computational models of reasoning with which TEC can be meaningfully contrasted (e.g., explanation-based learning programs like that of DeJong & Mooney, 1986; discourse processing models like that of Kintsch, 1988; and many others including those of Bar-on, 1991; Johnson, Krems, & Amra; 1994; Okada & Klahr, 1991; Ram & Leake, 1991; Shultz & Lepper, 1992; Thagard & Millgram, in press). For instance, Kintsch's (1988, 1992) Construction-Integration (CI) Model uses both rules and parallel constraint satisfaction (i.e., a "hybrid" model; cf. Smolensky, 1988) to comprehend a story. Using rules, CI constructs a weighted network of the text-based, inferred, and (randomly selected) associated propositions. (The text's propositions are given higher "preference" than inferred or associated ones.) The network is integrated by spreading activation until the system stabilizes—if it fails to stabilize, new constructions are added to the net and the integration process is repeated. Using this method, CI rejects "inappropriate" propositions. CI is similar to ECHO in that it uses parallel constraint satisfaction, and a kind of ECHO-like "data priority" for text-based propositions (over inferences and associates) in CI. Although ECHO doesn't deal with text comprehension per se, Thagard (1989) suggests that it can contribute to understanding and representing causal cohesiveness in stories.

Several other models of explanation evaluation and belief change are compatible with TEC and ECHO. For example, Ranney (in press) points out that TEC does not explicitly account for memorial capacity and processing limitations, inspiring Bar-On's (1991) theory of local coherence within *views*, an attempt to account for attentional and short term memory effects via limited capacities (cf. our version of WanderECHO; Hoadley et al., 1994, described in Chapter 2). Bar-On argues that localist connectionist models provide more appropriate levels of abstraction (than more distributed models) for simulating locally coherent views. Also similar to ECHO is HEIDER (Gabrys, 1989), which seeks consistency

(coherence) within its world view in the face of new information. Another related system is Shultz and Lepper's (1992) constraint satisfaction model of cognitive dissonance based on dissonance and consonance relations (analogous to contradiction and explanation in ECHO, although Shultz and Lepper randomly vary relation link weights within a given range across networks).

ECHO may initially seem less compatible with other computational models of reasoning, such as Ram and Leake (1991), Okada and Klahr (1991), and Johnson, Krems, and Amra (1994). For example, Ram and Leake (1991) argue that people prefer explanations that help them achieve their goals, and present a goal-based computational model that focuses on finding "useful" (vs. "valid") explanations by incorporating the goals into explanatory evaluations. Okada and Klahr (1991) code subjects' naive, complex, idiosyncratic, beliefs (garnered from transcribed protocols) as a hypothesis space, but they view belief revision as a search through this space of beliefs (vs. parallel constraint satisfaction, as in ECHO). However, both of these models highlight *goal-* or *utility-based* reasoning, which ECHO does not attempt to model. These models are more comparable to MOTIV-ECHO (Thagard, 1989) which allows ECHO's inferences to be biased by goals, or DECO (Thagard & Millgram, in press), which selects the most coherent plan given goals and actions (analogous to evidence and hypotheses).

Similar to Okada and Klahr (1991), Johnson, Krems, and Amra (1994) employ search through hypothesis (and experiment) space to model abduction in Abd-Soar. In Abd-Soar, they integrate and extend Soar (Laird, Newel, & Rosenbloom, 1987) and Klahr and Dunbar's (1988) model of scientific discovery as search through these dual spaces. Johnson et al. compare Abd-Soar to (somewhat anecdotal) subject data, and claim that unlike TEC (and ECHO), Abd-Soar models order and sequence effects (how the order of data presentation effects a response), as well as the power-law (practice-effect) speed-up. However, there are several

difficulties with their claims. For example, as the data and method were not detailed in Johnson et al., it is not clear that the participants actually believed one solution set over the other, as implied (e.g., were they asked or even given a chance to compare the two possible solutions described?). Also, order effects are not observed in all situations (e.g., Schank & Ranney, 1991). Further, in contrast to their claims, ECHO *can* model dynamic, sequential belief revision (e.g., Schank & Ranney, 1991 & 1992; Schank, Ranney, Mosmann & Montoya, 1993). ECHO can even come to different conclusions (i.e., land in different local minima) with different sequences (particularly as the coarseness of the simulation's stopping parameter is varied). Finally, any number of models can simulate the power-law speed-up; it's a property of the original model they chose to work with (Soar), as well as many other systems (e.g., ACT; Anderson, 1987), and not special to Abd-Soar. ECHO does not model power-law speedup because it is not a learning system, and does not generate hypotheses; it just evaluates them.

Related Prescriptive Work

Many researchers have illustrated difficulties that children and adults have with formal and informal reasoning (e.g., Chinn & Brewer, 1993; Kuhn, 1993; Linn & Songer, 1993; Markman, 1979; Perkins, Allen, & Hafner, 1983; Nickerson, Perkins, & Smith, 1985; Piaget, 1970; Schank & Ranney, 1992; Tversky & Kahneman, 1976). For example, Kuhn (1993) found that children and adults both tend to hold their theories with certainty and are often unlikely to (a) offer even simple evidence divergent from their theories, (b) comprehend evidence that would falsify their theories, or (c) develop or integrate counter-arguments. Markman (1979) shows a surprising insensitivity among elementary-school children to implicit inconsistencies, and sometimes even explicit inconsistencies. Chinn & Brewer (1993) show that undergraduates often ignore discordant information. Others (e.g.,

Case, 1974) have shown that specific reasoning strategies can be successfully taught to young children who lack the relevant formal reasoning strategies according to Piagetian theory (e.g., Inhelder & Piaget, 1958; Piaget, 1970). Given that many adults have difficulties using formal reasoning strategies, the development of formal reasoning strategies may not govern reasoning performance as suggested by Piaget—context and other factors may account for some of the variance (e.g., Linn, Clement, & Pulos, 1983). Perkins, Allen, and Haftner (1983) suggest that many people reason as though they are trying to minimize cognitive load rather than make sound inferences. Such "careless reasoners" casually elevate correlation to causation, affirm (vs. disconfirm) hypotheses, resist changing their hypotheses, and do whatever else keeps cognitive load low so long as the conclusion makes superficial sense and does not conflict with his or her intuitions. In contrast, ideal reasoners expend more effort to generate alternate hypotheses, and check or test claims in a variety of ways rather than relying on the weak heuristic of "intuitive fit."

Many others have discussed the need to teach reasoning skills, and have identified several useful skills to teach, or offered tools to support argument development (e.g., Bereiter & Scardamalia, 1989; Burbules, 1992, 1995; Friedler, Nachmias, & Songer, 1989; Giere, 1991; Hartley, Byard, & Mallen, 1991; Linn & Songer, 1993; Lipman, 1985, 1991; Markman, 1979; Nickerson, Perkins, & Smith, 1985; Perkins, 1985; Ranney, in press; Scardamalia & Bereiter, 1991; Smolensky, Fox, King, & Lewis, 1988; Toulmin, Rieke, & Janik, 1979; VanLehn, 1985). For instance, Markman (1979) argues that children need to be taught how to evaluate the plausibility of their own inferences, and suggests that teaching a few general evaluations principles (e.g., the plausibility of an inference increases with more coverage of observation) could improve their performance significantly. Nickerson et al. (1985) reviews methods and courses for teaching thinking skills, and identifies several useful skills to teach. Examples of such skills include: underlying reasoning

abilities (e.g., discriminating between hypotheses and evidence, hypothesis formation and evaluation), methods to support reasoning (e.g., the "scientific-method" heuristic, self-management strategies), knowledge about reasoning (e.g., general cognitive capabilities and limitations, common errors and biases in reasoning, one's own strengths and weaknesses), and fostering attitudes conducive to thinking (e.g., curiosity and wonder, the thrill of discovery, the satisfaction from productivity).

Lipman's (1985, 1991) "Philosophy for Children" curriculum teaches general thinking skills in (noncomputational) "communication environments" in which students learn to formulate questions and alternative hypotheses regarding philosophical issues. Toulmin, Rieke, and Janik (1979) present a general structure for arguments (extended from Toulmin, 1958), along with examples from science, ethics, and other fields. Computational frameworks for the construction and management of arguments (which use communication interfaces to reduce the ambiguity of arguments and to help reasoners record, modify, or invent arguments) have also been developed (e.g., Bereiter & Scardamalia, 1989; Carlock, 1990; Smolensky, Fox, King, & Lewis, 1988; VanLehn, 1985). The CSILE environment (Scardamalia & Bereiter, 1991) supports *group* dialectical processes by emphasizing questioning, idea generation and the sharing of information among students with the goal of building a group argument. Similarly, the Interactive Multimedia Kiosk (Hsi & Hoadley, 1994) supports collaborative knowledge building and reflection through the classification of comments into an argument map.

Attempts to teach Bayesian reasoning have had only limited success—not surprising given the difficulties people have with probabilistic reasoning, as mentioned above (e.g., biases in probability estimates and base rates; Tversky & Kahneman, 1976; Osherson, 1990). Beyond biases and calculation difficulties, reliable probabilities often aren't always available for most interesting real cases. Hence, most modeling of students' reasonings by probabilistic networks has been

conducted post-hoc (e.g., Sime, 1993; Villano, 1992). That is, experimenters usually (and laboriously) assign probabilities to a student's "reasoning dump" and model the result—suggesting little faith regarding students' ability to give on-line probability estimates, and making timely student feedback highly unlikely. These problems make instructional systems based on probabilistic models less feasible, helpful, and reliable than might be hoped.

Students often display a kind of "belief inertia" (Ranney, 1987/1988) in that they avoid resolving inconsistencies in general (cf. an imbalance toward assimilation vs. accommodation where "the characteristics of an object are not taken into account except insofar as they are consistent with the subject's momentary interests;" Piaget, 1970, p. 708). However, attempts to teach principles of coherent reasoning and stimulate belief revision in particular domains have shown some promise. For instance, Hartley, Byard, and Mallen (1991) describe a computer-based Newtonian modeling package that appears to help students build qualitative explanatory models of their understandings of physical motion by connecting causal links between objects and agents.

Research on the nature of students' beliefs about science suggest that students' epistemologies can affect how well they learn and integrate their understanding (e.g., Eylon & Linn, in press; Hammer, 1994; Linn & Songer, 1993). For instance, Linn and Songer (1993) categorized their students as having static, dynamic, or mixed beliefs about science (and categorized, respectively, 20%, 20%, and 60% of their students to have such beliefs). Indications of "static beliefs" included difficulty in discriminating between established and controversial ideas, and grouping all scientific assertions together as true; "dynamic beliefs" involved viewing science as a changing discipline; "mixed beliefs" indicated a combination of static and dynamic views. Linn and Songer looked at how such beliefs about science influenced students' test scores and understandings of physics—specifically, their ability to

understand isolated facts, integrate understandings, and propose principles. They found that students with dynamic beliefs of science improved their physics understanding the most as a result of instruction via case studies with highlighted principles. Eylon and Linn (in press), demonstrate that a computer learning environment that elicits student predictions and offers feedback regarding displaced volume seems to help students integrate their understanding; further, students with "cohesive" beliefs (e.g., who view scientific phenomena as governed by principles) sustained their understandings of displaced volume while students with "dissociated" beliefs (e.g., who view science as a collection of unrelated facts) did not. In this framework, *Convince Me* might also be a useful learning environment to help students understand scientific explanation and develop more cohesive/dynamic views of science.

Belvedere (Paolucci, Suthers, & Weiner, 1995; Suthers, Weiner, Connelly, & Paolucci, 1995) is a system developed to support individual and collaborative argumentation, based on Cavalli-Svorza's and others' work (Cavalli-Svorza, Lesgold, & Weiner, 1992; Cavalli-Svorza, Moore, & Suthers, 1993). The system offers a graphical interface to articulate arguments with views, although their several sorts of links (i.e., explains, supports, conflicts, causes, then, and) and argument objects (theory, hypothesis, claim, warrant, observation, drawing, text) may be opening a Pandora's box. That is, that their arguments are complicated to build and parse compared to the less baroque arguments generated with ECHO, with its two kinds of links (explanation and contradiction) and two basic kinds of argument objects (hypotheses and evidence). (A similar argument might be levied against the more complicated Toulminian arguments, which include backings, warrants, grounds, rebuttals, modalities, and claims; Toulmin, 1958). Their program does not provide model-based feedback (as *Convince Me* does, using ECHO), although the possibility of trying to incorporate a (perhaps even ECHO-like) "coach" has been discussed (e.g.,

Cavalli-Sforza et al, 1992). However, even if they were to use such a model, they would have many more decisions to make regarding setting link-weights, due to their large number of kinds of links. Thus, the model would have a high parameter/node ratio, and perhaps too many degrees of freedom to be useful.

In sum, while other systems exist for teaching reasoning and argumentation skills, *Convince Me* differs from these in that it is domain-general, is based upon a particular processing model, and it includes a computational model that actually yields predictions about the plausibility of an argument's assertions for the benefit of students—useful features that no other systems seem to have. The following chapters offer a summary of our prior descriptive studies of *Convince Me*'s underlying ECHO model, and describe the present *Convince Me* system and studies in detail.

2. THE GROUNDWORK: DESCRIPTIVE MODELING OF EXPLANATORY EVALUATIONS WITH ECHO

How well does ECHO model human reasoning? What are reliable methods for testing the model? Our research (e.g., Ranney & Schank, 1995) has traced a progression of methods for studying and aiding reasoning, which both assessed ECHO's modeling effectiveness and enhanced our understanding of the relationships among—and determinant features regarding—hypotheses, evidence, and the arguments that incorporate them. Unlike early (and others') work, which focused on *post-hoc* modeling of subjects' reasonings (e.g., regarding physical motion, social interactions, and juror reasoning; see below), we used ECHO to *predict*, a priori, the strength of subjects' beliefs. First, ECHO was used to predict subjects' text-based believability ratings (Schank & Ranney, 1991). Next, the bifurcation/bootstrapping method was developed to elicit and account for individuals' background knowledge (in this case, regarding pendular-release trajectories), while assessing inter-coder reliability regarding ECHO simulations (Schank & Ranney, 1992). Belief revision over time was also modeled, as were attentional and memory constraints (e.g., Ranney et al., 1993; Hoadley et al., 1994). Finally, the development of the *Convince Me* "reasoner's workbench" arose from the desires both to automate the explication of individuals' knowledge bases and their belief assessments, and to aid students in articulating and revising their theories (e.g., Schank & Ranney, 1993).

Post-hoc Modeling

ECHO has been successfully used to model juror reasoning (Thagard, 1989; also see Ranney et al., 1993) in which explanatory coherence plays a crucial role (Pennington & Hastie, 1988), to understand mental models of social interactions and relationships (Miller & Read, 1991; Ritter, 1991), and to model scientific reasoning (Thagard 1989, 1992). Ranney and Thagard (1988) presented the first application of ECHO to model on-line reasoning (e.g., in contrast to modeling arguments extracted from scientific treatises; Thagard, 1989, 1992; cf. Read & Marcus-Newhall, 1993), in which they simulated (*ex post facto*) changes both in subjects' beliefs and their conceptions of physical motion (Ranney 1987/1988; cf. Nersessian 1989). Using Ranney's (1987/1988) verbal protocols from subjects reasoning about ballistics, data derived from both rare and common conceptual difficulties were dynamically modeled. These subjects often achieved nontrivial Gestalt restructurings regarding inertia: For instance, one subject initially decided that objects dropped from a horizontally moving carrier (e.g., a train's window) would fall vertically, relative to the ground; she later realized, upon considering the apex of an arched trajectory, that such objects curve forward in their descents. Subjects' belief revisions were modeled in ECHO, which yielded activations that reasonably and temporally mimicked the beliefs that subjects rejected *or* accepted, as more information became available—either from subjects' personal inferences or from external sources (Ranney & Thagard, 1988).

Even dynamic post-hoc simulations of protocols, however, raise questions regarding the model's power relative to the data set's size (Ranney, in press). Hence, our later studies were designed to assess ECHO's *predictive* ability, as described below.

Predictive Modeling

Modeling Textually Embedded Propositions

Schank and Ranney (1991) examine three questions regarding ECHO's explanatory evaluations: Does ECHO predict differences in subjects' evaluations of different texts? Are local temporal order differences (not explicitly accounted for by ECHO) important to the subjects? Does ECHO predict inflectional reasoning, in which new information presented in texts (cf. Kintsch, 1988) yields significant changes in subjects' beliefs? In sum, the answers to these questions were Yes, No, and Yes, respectively. To investigate these questions, three contrasts between ECHO simulations and subjects' believabilities were conducted. Each contrast involved three stages: (1) ECHO simulations were run on given systems of propositions, (2) subjects read textual embodiments of the same propositional system and rated how strongly they believed each proposition, and (3) ECHO's ability to predict the subjects' believability ratings was assessed.

Undergraduate subjects read four fictional texts (counterbalanced for order) on the topics of medical diagnosis, wine tasting, linguistics, and physics. The texts reflected portions of the topology shown in Figure 2.1. For the first contrast ("Differential Predictions"), two topologies were used, the second of which includes an extra, critical piece of evidence (E4) that qualitatively changes the simulation results. Two groups of subjects were given texts reflecting these topologies (with one group given the extra piece of information). For the second contrast ("Temporal Order Effects"), two groups of subjects were given texts reflecting a single topology, with the second group given a text that provided the propositions in a different order. In the third contrast ("Dynamic Modeling"), subjects familiar with a text were given additional information (evidence E4, followed by hypothesis H3) that produced a "Gestalt switch" in ECHO. In all contrasts, subjects were asked to give believability

ratings for the textual statements, on a scale from 1 ("completely unbelievable") to 7 ("completely believable"), and their ratings were compared to ECHO's resulting activations for simulations run on the relevant topology.

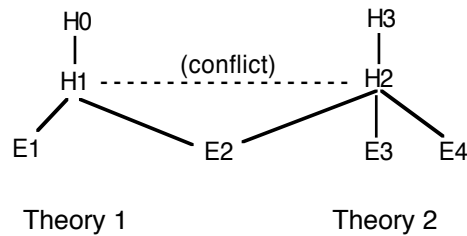


Figure 2.1. Topology of two conflicting explanatory theories {H0, H1, etc. vs. H2, H3, etc.}; dashes indicate theoretical conflict (from Schank & Ranney, 1991).

We found that subjects' believability ratings and inflectional reasoning were predicted well by ECHO's activations within a reasonable range of parameter values (with various parameter settings, overall $r = .67$ to $.74$, $p < .05$), and local temporal order differences did not significantly affect subjects' beliefs. Since ECHO does not automatically take such order differences into account, the results did not suggest necessary changes to the model. Further, our subjects often viewed competing hypotheses as non-exclusive, and indicated that they considered other information not present in the texts. This implicit information was modeled well in ECHO by giving certain hypotheses a fraction of data priority (again with various parameter values, overall $r = .77$ to $.78$, $p < .05$). However, this suggests that the subjects' representations of the situations were not completely captured by the representation encoded into ECHO.

These results inspired another study (described below; Schank & Ranney, 1992), in which we investigated ECHO's ability to model individual subjects' beliefs about physical motion when they made their "implied backings" explicit. A later study (Ranney et al., 1993) also assesses the utility of TEC's competition principle (given that subjects viewed apparently competing hypotheses as non-exclusive, unlike TEC), and ECHO's ability to model belief revision over time. In general, these later studies suggest that ECHO can model subjects' reasoning both on-line and over time, but that the current version of the competition principle adds no predictive utility to (and sometimes hinders) ECHO's modeling accuracy.

Modeling Verbal Protocols

Schank and Ranney (1992) explored a new, general method of assessing models that yield protocol-based predictions, so that these predictions can be contrasted with ratings data. In particular, our "bifurcation/bootstrapping" technique (see Figure 2.2) was applied to test how well ECHO could model on-line reasoning based on subject-generated arguments (as opposed to reasoning with largely text-constrained arguments—as in Schank & Ranney, 1991, and parts of Ranney et al., 1993). Intercoder agreement, a relative rarity among protocol-oriented analyses, is also readily assessed with this method (cf. Ericsson & Simon, 1993).

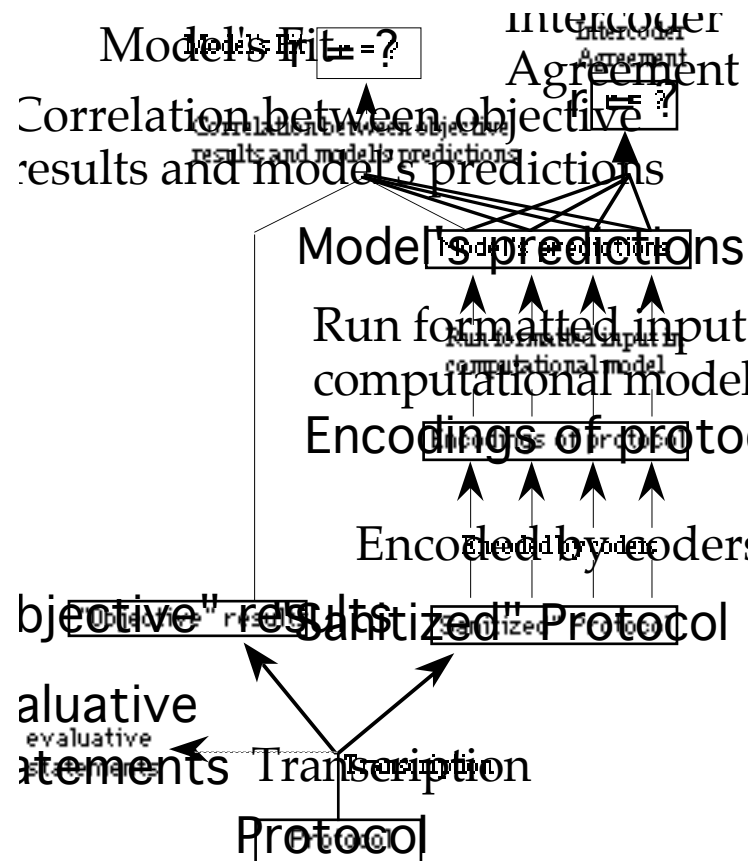


Figure 2.2. The "bifurcation/bootstrapping method" (from Schank & Ranney, 1992).

In this work, we initially asked subjects to predict (and explain) an endpoint pendular-release trajectory, while collecting believability ratings for their on-line beliefs (see Figure 2.3). Subjects were shown an animated pendular-release situation (from Ranney, 1987/1988), and as a subject reasoned out loud about the plausibility of her drawn (predicted) paths, the interviewer noted the subject's assertions. After the subject finished reasoning about the endpoint-release situation, the interviewer read back to the subject the list of beliefs she had noted. Subjects were then asked to rate (on a scale from 1, "completely unbelievable," to 9, "completely believable") how strongly they believed the propositions they had verbalized, and to rate how strongly they believed in their path. Five commonly predicted alternative trajectories

(from Ranney, 1987/1988, and a pilot study; see Figure 2.3) were then presented to the subject, and the process was repeated each time.

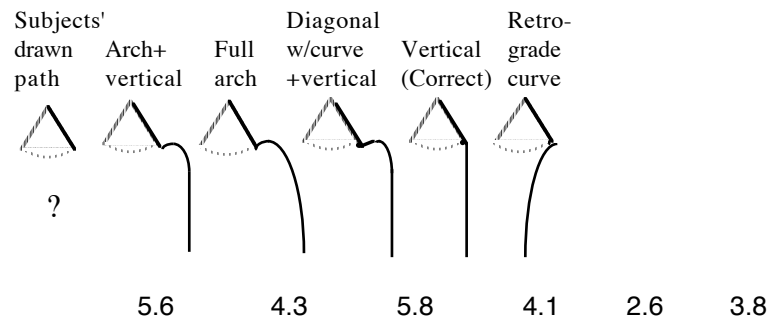


Figure 2.3. Mean pendular-release path believability ratings for the various alternatives, on a 1 to 9 scale, prior to feedback (from Schank & Ranney, 1992).

Subjects' stated believability ratings were edited out of copies of the transcribed protocols, as were evaluative statements that qualitatively revealed the strength of their beliefs. The edited protocols were then encoded into ECHO-style input by variously experienced, "blind" coders, who segmented and categorized subjects' assertions into beliefs, evidence, explanations, and contradictions. Simulations of the encodings were then run in ECHO, and comparisons between ECHO's activations and subjects' ratings (just prior to the time of feedback by the experimenter about the correct path) were made for each subject-coder pair and for each coder overall. To assess intercoder agreement, we examined the fit between ECHO's activations for coders' encodings of the same protocols.

We found that ECHO predicted subjects' ratings fairly well (overall $r = .56$, $p < .05$), although not quite as well as in Schank and Ranney (1991). The overall intercoder correlation ($r = .49$, $p < .001$) was no better than the overall model's fit

correlation , but both intercoder correlations and ECHO's predictive accuracy generally increased with coders' encoding experience. Analysis of variance results for the subjects' ratings over all beliefs later indicated that ECHO's activations explain about 28% of the variance in the subjects' ratings, while individual predilections account for about 8% of the variance in the ratings (both $p < .001$). As one would hope, the ANOVAs also indicate that *none* of the variance was accounted for by the coder, so systematic coder effects appear negligible. Still, we had hoped for higher intercoder correlations. This issue is obviated in later studies, since *Convince Me* is used to elicit subjects reasoning directly, eliminating the intercoder "middle person."

These data might suggest that ECHO does not predict subjects' beliefs better (or perhaps even as well as) when they make their implicit backings explicit. However, the task of modeling the subjects' beliefs in Schank and Ranney (1991) was of smaller scale—they were not encouraged to elaborate on their beliefs and bring other knowledge into their representations, as they were encouraged to do here. In addition, the ECHO networks generated here were, by salient measures (e.g., the number of propositions, the number of links), about two to over 20 times larger and much less explicit than the networks in the prior study. This extra complexity may have caused difficulties for subjects who, unlike ECHO, have limited attention and memory. These results inspired some extended modeling (described next) and the development of WanderECHO, to attempt to model attention and memory constraints.

Extended Dynamic Modeling of Protocols and Competing Beliefs

Ranney et al. (1993) presented two studies, analyzed with a threefold mission. First, we tested the predictive utility of TEC's auxiliary "competition principle," which suggests that people should infer an inhibitory relation between propositions

that *independently* explain another proposition (e.g., Thagard, 1992, 1991a). This principle had not yet been rigorously tested, but in many ways, it seems plausible: Suppose one hears that an evil dictator was dead due to a stabbing; then one later hears that he was dead due to gunshots. One might assume that the reports offer competing hypotheses (stabbing vs. shooting) for a datum (death). Given a situation in which two propositions each explain a third proposition (an explanandum), yet are not themselves explanatorily related, ECHO2 (with the competition principle) generates an incoherence link between the first two propositions by default. Results suggested that the competition principle needs refinement, that it probably overestimates subjects' abilities to infer and incorporate competitions among beliefs. Second, we assessed the dynamics of human belief revision as problems become more complicated. That is, as people reason about and articulate a problem more fully over time, does ECHO model them better? Or do they become overwhelmed by complexity and reason less coherently, as measured by ECHO? Results suggested that ECHO's fit increases over time for subject- (vs. experimenter-) generated arguments, and that subjects' local coherence, likely due to processing limitations, helps account for observed recency-related (and other) effects not modeled by ECHO. Third, we used these findings as a more informed foundation for using ECHO prescriptively, in *Convince Me*.

Materials used in these studies include (a) a problem relating to Berlin's location relative to the border that (until recently) divided Germany, and (b) pendular-release predictions from Schank and Ranney (1992), which were more molecularly and dynamically reanalyzed and remodeled here. Since the pendular-release task (b) is described in the previous section, I will focus here on the materials and method used in (a). We presented university undergraduates text segments that gradually biased them toward understanding Berlin's location. After each time segment, students rated the believability of every (isolated) proposition that had been read.

Results suggested that ECHO (without the competition principle) modeled students' belief evaluations better than its competition-principled counterpart, ECHO2. But this experiment may have represented an unfair test of the principle, given that only two (rather problematic) competitive links were simulated from the Berlin text. A more representative sample of competitive links might show their relative (if imperfect) descriptive power. In addition, while ECHO's modeling of this study's knowledgeable subjects (i.e., those who knew the location of Berlin a priori) improved over time, it *dropped* over time for naive subjects (i.e., those who did not know the location of Berlin a priori). This suggests that ECHO can lose descriptive power with increasingly complex reasoning—or it may have reflected some recency biases. Would this mean that naive students are bringing in other unmodeled information (other contexts), or reason less "rationally" (or in a less "globally coherent" fashion) with increased complexity? If the latter is the case, is it due to memory/attention limitations, or processing biased toward recent information?

We looked to the physics data for a more representative sample of competitive links, and another examination of ECHO's modeling over time. First, the encodings were parsed into seven accumulative segments, corresponding to periods during which the various sets of alternative trajectory predictions were considered and discussed by the students. These encodings were then used to run ECHO and ECHO2 simulations. Contrary to the first (Berlin) study's results, and for both variants, the rating-activation correlations generally show an *improving* fit over time—even though these physics data were considerably more complex than the Berlin data. The simulations also suggested that the competition principle may be useful, albeit imperfect; for instance, 57% of the competitive links generated by ECHO2 were judged to truly capture an underlying competition. Further, analyses of students' arguments revealed that subjects were biased toward lingering and elaborating on the most recently generated regions of their argument.

In sum, both studies indicate that TEC's competition principle seems to need further revision since it lends little or no predictive utility. These results further inspired the development of WanderECHO to model attentional and memorial constraints, as described below. Finally, we also hypothesized that *Convince Me* would help students overcome the inherent processing limitations (e.g., unassisted short-term memory) and biases (e.g., recency effects and locally coherent reasoning) observed here: *Convince Me* provides both memorial support for argument development via its argument interface, and globally-coherent feedback that is unbiased by the order of information. This feedback lets students know whether the strengths of their beliefs are in accord with their argument structures, and focuses them on where they and ECHO most disagree. Highlighting such disparities might help students pinpoint possible inconsistencies in their arguments, encouraging reflection and revision.

Modeling Attention and Memory Constraints with WanderECHO

From insights arising from our descriptive studies of ECHO, we developed a variation of the model, WanderECHO, that attempts to simulate a traveling focus of attention (Hoadley et al., 1994). Several variants of the WanderECHO simulation were applied to Schank and Ranney's (1991) data, and were found to generally simulate subjects' mean believability ratings better than "standard" ECHO.

One might argue that ECHO is flawed in its representation of human thought in at least two aspects: computational power and memory. That is, the model does not account for human limitations; ECHO will run on a network of almost unlimited size and will continue updating all activations until the entire network virtually settles. WanderECHO is a variation on ECHO that tries to take some of these considerations into account. First, WanderECHO simulates a limited focus of

attention and does not demand massively parallel execution. At any given cycle, rather than updating every unit in the network, it updates one unit; this node is the model's "focus of attention." The first node to be updated is chosen randomly; the next node to be updated is chosen probabilistically based on link weights. Second, WanderECHO has a local stopping criterion, and does not require calculation of the energy change of the entire system in order to determine whether or not to stop; it will satisfice, rather than optimize as ECHO does.

Since the WanderECHO simulation is stochastic, comparing human results to a single run of the model is inappropriate. So, the model was run 200 times and output activations were averaged across them. These average activations were then correlated with the data from Schank and Ranney (1991) on believability ratings of textually embedded propositions. Most of the activation-rating correlations were numerically greater than that of the "best-parameters" ECHO simulation (Schank & Ranney, 1991), and all modeled the data significantly better than ECHO with Ranney and Thagard's (1988) default parameters. While the data set is too small to be conclusive, the results are encouraging regarding WanderECHO's prospects for becoming a useful simulation of limited coherence.

Automating Knowledge Elicitation and Supporting Argument Development with *Convince Me*

Even though our protocol modeling results (e.g., Schank & Ranney, 1992) indicate both reasonably good data-fitting and inter-coder reliability, the bifurcation/bootstrapping method is fairly unwieldy in that it requires an extremely vigilant and well-practiced experimenter. Also, certain Gricean maxims of conversation—and normal information processing limitations on the part of the

experimental interviewer—meant that the record of argumentation would always be somewhat spotty when derived via the bifurcation/bootstrapping method (see, e.g., Grice, 1975, on the *maxim of quantity*). Comparable data can now be recorded in a more automated, yet rigorous, fashion, using *Convince Me* (e.g., Schank & Ranney, 1993), which captures both (a) a subject's "knowledge dump" of evidence and hypotheses—including their relationships, and (b) believability ratings for the proposed beliefs. Rather than eliciting these in the maelstrom of an on-line interview/protocol session, *Convince Me* can function as a "reasoner's workbench" with which subjects explicate their beliefs about a controversy. Further, we expected that *Convince Me* would help students generate more coherent arguments by reducing processing limitations and biases (as discussed above), encouraging students to explicate their reasoning, focusing them on discrepancies in their arguments, and encouraging reflection and revision. The design and instructional effectiveness of the system are the focus of the following chapters.

3. *CONVINCE ME*

Convince Me is a computational "reasoner's workbench" program that supports argument development and revision, and provides ECHO-based feedback on the coherence of subjects' articulated beliefs. The associated curriculum discusses distinctions between hypotheses and evidence, strategies for generating and evaluating (everyday and scientific) arguments and counter-arguments, and reasoning biases and how to reduce them. *Convince Me* is a domain-independent system in that it is as applicable to theoretical debates in biology (as in the BioQUEST library; Schank, Ranney, & Hoadley, 1995) as it is to wine-tasting (as in Schank & Ranney, 1991). It was implemented by the author in HyperCard (with external C commands), and runs on a Macintosh with a 13" (Powerbook-size) or 17" (two-page) monitor (Schank, Ranney, & Hoadley, 1995).

Using *Convince Me*, students can (a) articulate their beliefs regarding a controversy, (b) categorize each notion as either being evidential or hypothetical, (c) connect their beliefs inhibitorily and/or explanatorily, (d) provide ratings to indicate the believability of each statement, and (e) run the ECHO simulation to obtain various forms of feedback. *Convince Me* also incorporates various other important features, including the ability for students to continually modify their arguments, belief ratings—and even the parameters that govern ECHO's "reasoning engine" (see the description of ECHO in the Introduction). The main features of the system are detailed in an example session below.

An Example Argument

Consider the following situation (one of many in the curriculum), which involves two different viewpoints about a common situation—freezing ice cubes:

"Latisha's Mom says that to make ice cubes freeze faster, you should use hot water instead of cold water in the ice cube tray (H1). She has been doing this for many years, and although she didn't believe it when she first heard it, Latisha's Mom tried it out several times and the hot water did freeze faster (E2).

Latisha learned in science class that it takes longer for hot things to cool to room temperature than it takes for warm things which are closer to room temperature (E1). She thinks that water freezing should behave the same way as objects cooling to room temperature (H3), which suggests that cold water would freeze faster (H2)."

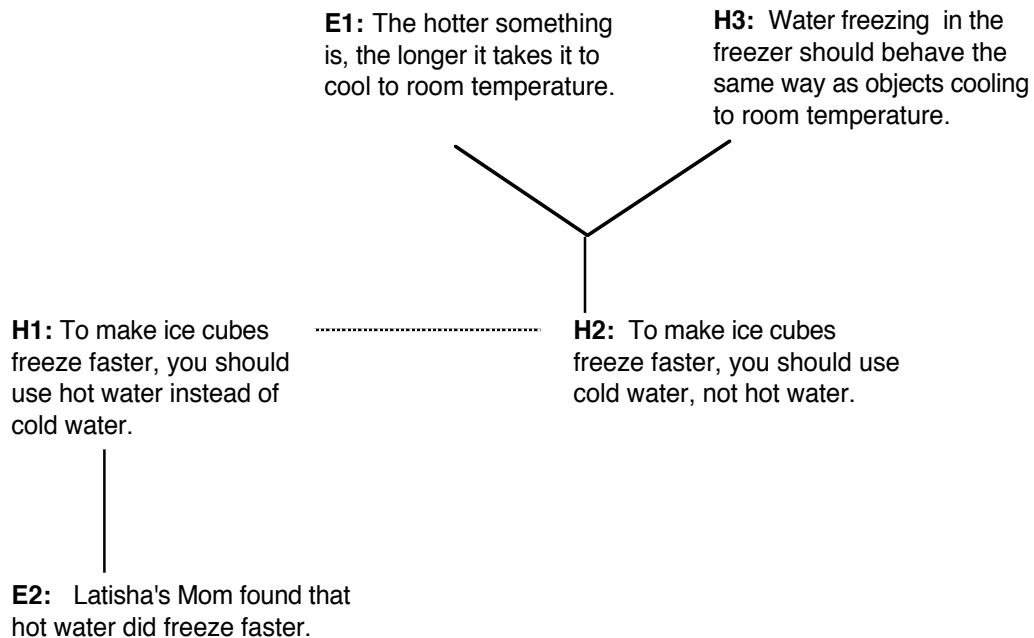


Figure 3.1. Two viewpoints about freezing ice cubes (sample topology).

Figure 3.1 shows a sample topology derived from this text. Given such a situation, the student can use *Convince Me* to enter her ideas. After adding (or editing) a statement, she is asked to (a) check any number of four phrases that apply ("Acknowledged fact or statistic," "Observation or memory," "One possible inference, opinion, or view," "Some reasonable people might disagree") in order to help determine if the statement is a hypothesis or a piece of evidence, (b) to explicitly choose one of the two (evidence/hypothesis) categories, and (c) to specify the reliability of beliefs she classifies as evidence (see the bottom dialog box in Figure 3.2). The student can also indicate which ideas explain (either independently or jointly) and contradict which other ideas (see Figure 3.3 and 3.4).

After entering an entire argument, the student is asked to rate how strongly she believes each statement (see Figure 3.5), and then run an ECHO simulation to see which statements her argument helped to support or reject and which ones it left neutral—from the simulation's point of view. After the simulation has run, small "thermometer" icons show up on the screen, one for each statement (see Figure 3.2, upper right; cf. the diagrammatic interface in Chapter 5, "Future Directions"). The higher the mercury, the more ECHO accepts the statement; the lower the mercury, the more ECHO rejects the statement. Numerical equivalents of these iconic measures also appear beside the ratings provided by the student, who can then compare her ratings with ECHO's output, statement by statement (see Figure 3.2, upper middle). In addition, she can also ask *Convince Me* to report (a) a "model's fit" correlation between her ratings and ECHO's scaled activation values, (b) how related the two sets of ratings are (e.g., "mildly opposed", "moderately related", "highly related"), and (c) which (three) pairs of values differ the most (see the center box in Figure 3.2).

CM (old, no graphing)

Statements: (Add...) (Edit...) (Delete) (Rate...) (Rate All...) (Model's fit...)

Hypotheses:

Hypothesis	Rating	ECHO
H1. To make ice cubes freeze faster, use hot water, not cold water	5	6.7
H2. To make ice cubes freeze faster, use cold water, not hot water	6	4
H3. Water in the freezer should behave the same way as objects cooling	7	5.5
H4. More of the hot water evaporates so there's less mass to freeze		

Evidence:

Evidence	Rating	ECHO
E1. The hotter something is, the longer it takes it to cool to room temperature	8	7.3
E2. Latisha's Mom found that hot water did freeze faster	7	7

Explanations: (Explain...) (Explain All...) (Delete Explanation)

The statement(s) that explain(s) "H2. To make ice cubes freeze faster, use cold water, not hot water" is/are:

H3. Water in the freezer should behave the same way as objects cooling to room temperature "AND"
E1. The hotter something is, the longer it takes it to cool to room temperature

Contradictions: (Conflict...) (Conflict All...) (Delete Conflict)

The statement(s) that conflict(s) with "H2. To make ice cubes freeze faster, use cold water, not hot water" is/are:

H1. To make ice cubes freeze faster, use hot water, not cold water

Simulation results:

Hypotheses:

H1(6.7) H2(4) H3(5.5)

Evidence:

E1(7.3) E2(7)

Oops! (undo)

HelpMessages:

E1 The hotter something is, the longer it takes it to cool to room temperature

Steps for using CONVINC ME:

1. Enter hypotheses and evidence.
2. Enter explanations and contradictions.
3. Rate the believability of your statements.
4. Run the ECHO simulation.
5. Compare your evaluations to ECHO's.
6. (optional) Make changes based on ECHO's feedback.

Current File:

The correlation between your ratings and ECHO's evaluations is: 0.34 (mildly related).

The three most disparately rated statements are: H2, H1, H3, respectively (see bolded statements).

Your statement:

More of the hot water evaporates so there's less mass to freeze

Check all that apply:

Acknowledged fact or statistic

Observation or memory

One possible inference, opinion, or view

Some reasonable people might disagree

Select one:

Evidence E3 Reliability, if evidence? (from 1, poor, to 3, good)

Hypothesis H4

OK Cancel


Figure 3.2. A user adds and classifies a belief about the speeds at which water of different initial temperatures freeze (bottom) in response to *Convince Me's* feedback (middle). (Cf. the diagrammatic interface in Chapter 5, "Future Directions.")

What (if anything) explains the statement:
H1. To make ice cubes freeze faster, use hot water, not cold water

(Use command-click to select more than one statement.)

H2. To make ice cubes freeze faster, use cold water, not hot water
H3. Water in the freezer should behave the same way as objects cooling to room tem...
H4. More of the hot water evaporates so there's less mass to freeze
E1. The hotter something is, the longer it takes it to cool to room temperature
E2. Latisha's Mom found that hot water did freeze faster

Choose one:

Each statement explains the claim **independently** 

Statements **jointly** explain the claim

Figure 3.3. Adding an explanation.

What (if anything) conflicts with the
H1. To make ice cubes freeze faster, use hot water, not cold water

(Use command-click to select more than one statement.)

H2. To make ice cubes freeze faster, use cold water, not hot water
H3. Water in the freezer should behave the same way as objects cooling to room tempe...
H4. More of the hot water evaporates so there's less mass to freeze
E1. The hotter something is, the longer it takes it to cool to room temperature
E2. Latisha's Mom found that hot water did freeze faster




Figure 3.4. Adding a contradiction.

How strongly to you believe the statement:

H4. More of the hot water evaporates so there's less mass to freeze

On a scale from 1 (completely disbelieved/false) to 9 (completely believed/true)?

Use All Old Ratings

6

OK Cancel

Figure 3.5. Rating a statement's believability.

Parameters: Use Default

Explanation:	<u>0.03</u>	↑ ↓
Conflict:	<u>0.06</u>	↑ ↓
Data 'boost':	<u>0.055</u>	↑ ↓
Skepticism:	<u>0.04</u>	↑ ↓

Figure 3.6. Modifying ECHO's parameters (default values are shown).

Based on *Convince Me's* feedback, the student can modify her ratings and/or the structure of her argument (perhaps focusing on the statements regarding which she and ECHO most "disagree"). Users are even permitted to alter the ECHO model if they feel that it doesn't "reason as they do." Figure 3.6 shows how a user may change the levels of "skepticism" (activation decay), data priority (or 'data boost'), and the relative importance of explanations (excitation) and "conflicts" (inhibitory

contradictions and competitions). However, users rarely find it necessary to question ECHO's default parameters, as they usually prefer to further explicate their arguments first.

Essentially, ECHO's feedback lets students know whether their beliefs are in concert with the coherence of the argument that they articulated and entered into *Convince Me*. Thus, *Convince Me* seems to be the only working system that both assists the elucidation of students' thinking while providing them with simulation-based feedback about the coherence of their articulated beliefs and mental representations. The particular variant of the ECHO simulation employed in *Convince Me* also represents an advance over that employed by Ranney and Thagard (1988) and others, in that pieces of evidence are differentially weighted by how reliable the users say they are (which they indicate when they categorize them as evidence). This is important, as evidence can vary in believability as a function of the various methods that spawn it (e.g., Ranney, in press). In fact, a piece of evidence might be considered a place-holder for an entire "subnetwork" argument regarding an observation and its methodological context.

Associated Materials

The associated materials essentially comprise about 3 hours worth of tests and 5 hours of intervention and exercises (8 hours in total). In order of intended use (see Figure 3.7), this includes a pre-test (approximately 90 minutes), three curriculum units on scientific reasoning (approximately 1 hour each), integrative exercises (which may be used with or without the *Convince Me* software; approximately 2 hours), a post-test (which replaces some pre-test items with new isomorphic items; also approximately 90 minutes), and an exit questionnaire (approximately 10 minutes). All materials are given in Appendices A-G.

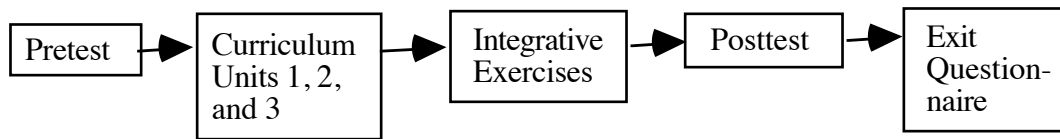


Figure 3.7. Associated materials and the intended sequence of use.

Pre-Test

The 90-minute paper pre-test assesses reasoning skills such as one's ability to classify hypotheses and evidence, evaluate scientific theories, and generate and disconfirm alternate hypotheses (e.g., using tasks that include some related to those of Wason, 1968, and Wason & Johnson-Laird, 1972). First, participants are asked to provide definitions for hypothesis, evidence, fact, explanation, contradiction, theory, argument, confirmation bias, disconfirmation, recency bias, and primacy bias. Next, they rate several statements presented in isolation (see Table 3.1) or within a story context (see Table 3.2) on a 1-9 scale—in terms of their believability, and as exemplars of hypothesis and evidence (i.e., in the way that one might search for prototypical hypotheses and evidence; cf. Rosch, 1977). Then they are asked to generate hypotheses, attempt disconfirmations, and offer data regarding two given situations. Finally, for two given passages, they identify (and give believability ratings for) hypotheses and evidence in each passage, propose (and rate) alternative propositions, and state which propositions explain and contradict which others; subjects are also prompted to make any revisions to their argument and ratings, as they desire. (The pre-test is shown in Appendix A.)

Table 3.1. Rating instructions and examples of isolated propositions.

 Rating instructions:

Based on **your** view and knowledge of the world, for each of the following statements please:

1. Rate (circle) how good an example of a *hypothesis* you think the statement is,
2. Rate (circle) how good an example of a piece of *evidence* you think the statement is,
3. Explain (briefly, in writing) why you gave the hypothesis and evidence ratings you did, and
4. Rate (circle) how strongly you *believe* the statement.

Some examples of the isolated propositions, available for rating:

All wine is made from grapes.
 Gravity exists in other galaxies.
 President John F. Kennedy was assassinated.
 Abraham Lincoln said that Ross Perot would lose in 1992.
 Birds evolved from animals that lived in trees.

Propositional rating example:

a) President John F. Kennedy was assassinated.

definitely <u>not</u> hypothesis	neutral	definitely hypothesis
1 2 3 4 5 6 7 8 9		

definitely <u>not</u> evidence	neutral	definitely evidence
1 2 3 4 5 6 7 8 9		

why (explain): _____

completely disbelieve/reject	neutral	completely believe/accept
1 2 3 4 5 6 7 8 9		

Table 3.2. Propositions embedded within a story context.

Some dogs have an aggressive disorder. They bark more than other dogs, growl at strangers, and sometimes even bite. They also tend to have higher blood pressure and heart rate than other dogs.

Some researchers think that these dogs get the aggressive disorder when their owners treat them poorly, that is, when the owner neglects the dog, doesn't give it enough love, or hits it. These researchers trained one group of aggressive-disorder dog owners to treat their dogs firmly yet lovingly. They found that all dogs whose owners were trained barked much less, were much friendlier to strangers, never bit a stranger, and had lower heart rate and blood pressure than dogs whose owners had not been trained. These researchers said that their experiment proved that abuse causes dogs to have the disorder.

Other researchers disagree. They think that dogs with the disorder are born without a certain chemical in their body. They think that the lack of this chemical elevates their blood pressure and causes the disorder. These researchers gave one group of aggressive-disorder dogs a medicine that contained the chemical. They found that the dogs had a much lower heart rate and blood pressure, were friendlier to strangers, did not bark as much, and never bit anyone. These researchers said that their experiment proved that the missing chemical causes dogs to have the disorder.

Examples of propositions from the above context, made available for rating:

Some dogs have an aggressive disorder.

Some researchers think dogs get an aggressive disorder when their owners treat them poorly.

Abuse causes an aggressive disorder in dogs.

Some researchers found that a chemical relieved symptoms of aggressive disorder in dogs.

Curriculum Units

The three-hour paper curriculum includes one unit on evidence, hypotheses, and theories, a second unit on reasoning about arguments, and a third unit on how to use *Convince Me*. The curriculum units are included in Appendices B, C, and D.

Unit 1, "Evidence, Hypotheses, and Theories"

Unit 1 is designed to help students (a) think about the hypothesis-evidence continuum, and identify and justify distinctions between hypotheses and evidence, (b)

generate and classify hypotheses and evidence, relate them (via independent or joint explanatory relations, and/or contradictory or neutral relations) to create an "explanatory theory," and justify classifications and relationships, (c) evaluate the believability of evidence and hypotheses, and justify evaluations, and (d) identify and reconcile contradictions and competing (alternative) explanations. (See Appendix B.)

Unit 2, "Reasoning About Arguments"

This unit is primarily designed to help students (a) understand the need for alternative hypotheses (e.g., to overcome confirmation bias⁵), (b) generate complete arguments based on given scientific or everyday controversies, (c) reduce confirmation, primacy, and recency biases (e.g., as observed in Ranney et al., 1993), and form and change their opinions about an argument. (See Appendix C.)

Unit 3, "Using *Convince Me*"

Unit 3 describes how to use this software to enter, save, and evaluate arguments. In particular, it explains to students how to (1) input their own situational beliefs, (2) classify them as hypotheses or evidence, (3) indicate which beliefs explain or contradict which others, (4) rate their beliefs' plausibilities, (5) run the ECHO simulation, (6) contrast their ratings with ECHO's predictions, and (7) modify ECHO's parameters to better model their individual reasoning style, if they

⁵Attempting to confirm a hypothesis doesn't necessarily indicate an irrational confirmation bias. Rather, as Holyoak and Spellman (1993) argue, successful hypothesis testing "...often involves an initial focus on confirmation followed by more critical examination of 'loose ends' or apparent anomalies, which may lead to hypothesis revision." That is, focus on confirmation and disconfirmation can (rationally) vary dynamically over the course of an inquiry.

so desire. Throughout the curriculum, participants are also encouraged to modify their arguments or ratings as needed. (See Appendix D.)

Integrative Exercises

After completing Unit 3, students are given a set of four integrative exercises. For each exercise, students are given (a) a passage presenting two competing theories, and (b) a set of instructions. The passages, in the order presented, include competing theories regarding animal behavior (specifically, yawning), medical diagnosis, expected pendular release trajectories (from Schank & Ranney, 1992), and views on abortion. The passages generally decrease in the length and detail given. The instructions for *Convince Me* users are to enter arguments and believability ratings on these topics into the system and to run the ECHO simulation, making revisions as they wish. The exercises can also be completed on paper (without the software) by simply listing the hypotheses, evidence, and relations between them, as well as the believability ratings and (any later) modifications. The passages and instructions for completing the exercises (with or without the software) are given in Appendix E.

Post-Test

The 90-minute paper post-test is similar to the pre-test, and again assesses one's ability to classify hypotheses and evidence, generate and disconfirm alternate hypotheses, and evaluate scientific theories (see "Pre-test" description above, and Appendices A and F). Three sets of items on the post-test were identical to those on the pre-test: the definitions, and both the isolated and contextualized statements available for rating. The remaining four sets of items were isomorphic to those on the pre-test (i.e., those involving hypothesis generation and disconfirmation, as well as the identification of evidence, hypotheses, explanations, and contradictions from a given passage).

Exit Questionnaire

The exit questionnaire asks students to (a) rate and describe how much they learned from the software, exercises, tests, and each of the curriculum units, and (b) describe what they liked most and least about the software, exercises, and curriculum—and offer any suggestions for how to improve them. Students are given copies of the curriculum units to refer to while completing the questionnaire. The questionnaire takes approximately 10 minutes to complete. (See Appendix G.)

The following chapter discusses two prescriptive studies conducted with the *Convince Me* software, curriculum, and test materials that were described here.

4. PRESCRIPTIVE STUDIES USING *CONVINCE ME*

Two prescriptive studies with *Convince Me* were conducted, and are described here. The focus of the first study was on the utility of the *Convince Me* software and curriculum, as well as the nature of both novice and expert notions of evidence, hypothesis, and related constructs regarding critical thinking (e.g., Ranney, Schank, Hoadley, & Neff, 1994). Results suggest that, while the distinguishing characteristics of data and theory are still vague—even for experts—the system lends a sophistication to novices' discriminative criteria, making their epistemic categorizations seem more expert-like. The second prescriptive study addressed the question of whether the system is a tool and/or a training device to yield more coherent argumentation skills. That is, does *Convince Me* make its users (a) better reasoners while they employ it, (b) better reasoners even when they are distal from it, (c) both, or (d) neither? The empirical results indicate that *Convince Me* is useful tool that even yields transfer to unsupported practice.

Study 1: Experts vs. Novices, and The Hypothesis/Evidence Distinction

Researchers commonly suggest that it is desirable for children and lay people (and perhaps even some scientists) to improve their understanding of the evidence/hypothesis distinction (e.g., Kuhn, 1989). Most (and especially empirical) researchers—including ourselves—have implied that the distinction is either easy to make, or that at least skilled scientists make it fairly well (cf. Giere, 1991). Definitions in science books often suggest the former, as if context does not have a major impact on the epistemic categorization. But as discussed in Chapter 1 (see

"Philosophy," under "Related Descriptive Work"), others argue that the observation/theory classification is not clear-cut (e.g., Feyerabend, 1978; Hanson, 1958/1965; Longino, 1990).

Little controlled experimental work on this issue, aside from that offered here, exists. The present study draws upon (a) our past empirical insights into how syntax and word choice can bias the evidence/hypothesis classification (e.g., Schank & Ranney, 1991) and (b) issues addressed while designing the *Convince Me* "reasoner's workbench" software, which considers evidence to be variably reliable—and hence variably worthy of the full computational effects of data priority (see Ranney, in press; Schank & Ranney, 1993, etc.). Several questions regarding the hypothesis/evidence distinction were considered: First, how are particular individual propositions classified? This is addressed by observing how participants rate the "hypothesis-likeness," "evidence-likeness," and "believability" of a corpus of scientific statements. Second, how does context seem to affect these classifications? To address this, statements were provided either in isolation or within a textual, story-type, context. Third, how do experts in scientific reasoning differ from untrained novices in classifying these statements? To investigate this, samples of the two populations were compared, particularly regarding their inter-construct relationships and inter-subject agreement. Finally, how accurate and useful are definitions of "hypothesis," "theory," "evidence," "fact," and other such constructs? For this question, novices were asked to define the set of terms, and experts were asked to grade the goodness/accuracy of these novices' definitions.

One might expect that average ratings of evidence-likeness and hypothesis-likeness are (or should be) negatively correlated; this distinction would reflect differences in their relative controversy, contestability, reliability, and perceptibility. Further, since ECHO's data priority should lend activation more to evidence than to hypotheses, believability should also be—again, on average—negatively correlated

with hypotheses, while positively correlated with evidence. Another reason for expecting negative correlations involving hypotheses stems from the many situations in which one has more than two (perhaps implicit) alternate hypotheses that cover the same scope (e.g., several competing ideas about dinosaur extinction), although only one is likely to be correct. Hence, most of one's hypotheses are likely to have low believability, while most of one's evidential propositions should have higher believability, due to data priority and the relatively fewer inhibitory relations—such as competitions and contradictions—associated with evidence. This pattern is certainly in concert with what we have observed in past studies (e.g., Schank & Ranney, 1991; see "Discussion" below).

Method

Participants

Ten novices and ten experts participated in this study. The novices (four women and six men) were undergraduate students from the University of California, Berkeley. (The terms "novice" and "student" are used interchangeably throughout this study.) They responded to campus advertisements and were paid five dollars per hour for their participation. Their backgrounds were varied, but they had essentially no background in logic or the philosophy of science. The expert volunteers were from the University of Chicago, the University of California (Berkeley), Princeton University, the Tennessee Institute of Technology, and the Educational Testing Service. (Five were post-Ph.D., and five were doctoral students; three were women, and seven were men.) The experts had experience in cognitive science, the philosophy of science, science education, and logic, and each is currently studying scientific and practical reasoning. Nine experts provided propositional (statement) ratings, and five experts provided goodness ratings of novices' scientific definitions. (Four experts provided both propositional and definitional ratings.)

Design and Procedure

As shown in Figure 4.1, the novices completed the pre-test, three curriculum units on scientific reasoning, integrative exercises using *Convince Me*, the post-test, and exit questionnaire, as described in Chapter 3, with one exception—the "dogs" statements (#2 b, d, f, h, j, l, n, p; see Appendices A & F) were not included as *isolated* statements on the tests (as they were in Study 2), but *were* included in a story context (see problem 4, Appendices A & F). One subgroup of ("propositional") experts was asked to complete the proposition-rating portions of the pre-test. The other subgroup of ("definitional") experts was given a randomly-ordered booklet of novices' completed definitions from the pre- and post-tests, and were asked to score them on a scale from 1 (poor) to 3 (good) for each given definition.

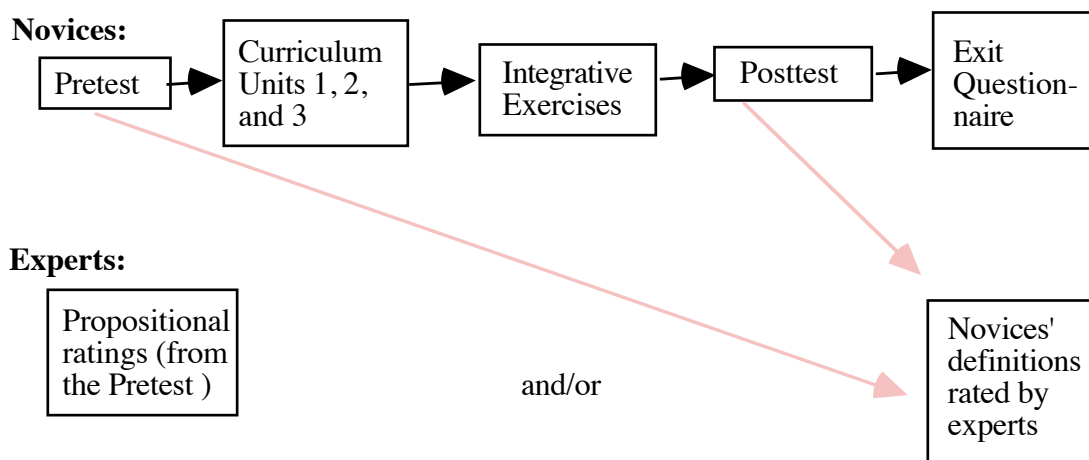


Figure 4.1. Summary of this experiment's method.

Results

Two-tailed tests at the $p=.05$ level were conducted for all analyses, unless otherwise noted. As shown below, context and training generally improved novices' hypothesis/evidence classifications, yet even experts found them difficult:

Propositional Ratings

Correlations among the constructs of evidence, hypothesis, and believability.

As Table 4.1 illustrates, context generally adds to the discriminability between evidence and hypotheses across groups and times of testing. Even experts, who exhibited a statistically significant negative correlation ($-.28$) for no-context propositions, improved the magnitude of their evidence-hypothesis distinction in context to $r = -.66$ (all p 's $< .05$ unless otherwise noted).

Without a context, novices initially show no significant correlation between evidence and hypothesis ($r = -.03$), but training ($-.30$), context ($-.41$), and both factors together ($-.63$) significantly increase the absolute value of the observed relationships. The novices also showed a similar pattern of results with respect to their believability-hypothesis distinction, with a nonsignificantly positive correlation ($r = .09$) becoming highly and significantly negative ($-.68$) due to context and training with *Convince Me*. Furthermore, training played a role in significantly increasing the novices' initial (and significant) in-context believability-evidence correlation from $.42$ to $.64$.

Training generally made novices behave more like experts. Experts exhibited negative evidence-hypothesis correlations ($-.28$ out of context and $-.66$ in context), and novices achieved these levels during post-testing ($-.30$ out of context and $-.63$ in context, vs. $-.03$ and $-.30$ during their pre-test). Further, novices eventually approximated the experts' negative believability-hypothesis correlation for no-context propositions ($-.14$ vs. $-.24$). While novices' believability-hypothesis correlations

were more negative than experts' in context (-.57 and -.68 vs. -.19), in general, both groups had fairly positivistic stances, as believability-evidence correlations ranged from .35 to .64 across participants, context-types, and testing times. (Note that, after training, novices exhibited the largest of the believability-evidence correlations.)

Table 4.1. Within-group correlations between believability and hypothesis-likeness (B-H), evidence-likeness and hypothesis-likeness (E-H), and believability and evidence-likeness (B-E), Study 1 (from Ranney et al., 1994).

		Novices			Experts		
		<u>B-H</u>	<u>E-H</u>	<u>B-E</u>	<u>B-H</u>	<u>E-H</u>	<u>B-E</u>
No context*:	pre	.09	-.03	.48 ^a	-.24 ^{ab}	-.28 ^a	.35 ^a
	post	-.14	-.30 ^a	.60 ^a			
In context:	pre	-.57 ^{ac}	-.41 ^{ac}	.42 ^a	-.19 ^b	-.66 ^{abc}	.37 ^a
	post	-.68 ^{ac}	-.63 ^{abc}	.64 ^{ab}			

*Isolated "dogs" statements (#2 b, d, f, h, j, l, n, p) not included.

^a $r \neq 0$, $p < .05$, 2-tail $Z = 1.96$

^bsignificantly different from novice's pretest, $p < .05$, 2-tail $Z = 1.96$

^csignificantly different from no-context, $p < .05$, 2-tail $Z = 1.96$

Table 4.2. Between-group correlations regarding believability (B-B), evidence-likeness (E-E), and hypothesis-likeness (H-H), Study 1 (from Ranney et al., 1994).

		Novices			Experts		
		<u>B-B</u>	<u>E-E</u>	<u>H-H</u>	<u>B-B</u>	<u>E-E</u>	<u>H-H</u>
No context*:	pre	.66 ^a	.31 ^a	.15 ^a	.87 ^{ab}	.20 ^a	.28 ^a
	post	.65 ^a	.32 ^a	.06			
In context:	pre	.20 ^{ac}	.23 ^a	.29 ^{ac}	-.04 ^{bc}	.42 ^{abc}	.54 ^{abc}
	post	.25 ^{ac}	.44 ^{ab}	.39 ^{ac}			

*Isolated "dogs" statements (#2 b, d, f, h, j, l, n, p) not included.

^a $r \neq 0$, $p < .05$, 2-tail $Z = 1.96$

^bsignificantly different from novice's pretest, $p < .05$, 2-tail $Z = 1.96$

^csignificantly different from no-context, $p < .05$, 2-tail $Z = 1.96$

Inter-rater agreement regarding the constructs of evidence, hypothesis, and believability. As shown in Table 4.2, both groups showed greater inter-rater reliability (correlations) across their believability ratings for the no-context propositions, regardless of testing time. This is not surprising, since the in-context situation (shown in Table 3.2), involving the age-old nature-nurture issue, is a particularly controversial one (i.e., of low systemic coherence; Schank & Ranney, 1992) compared to the less subtle no-context items. In contrast, there was *less* agreement regarding the hypothesis-likeness of no-context propositions (relative to in-context propositions), and effects in the same direction regarding the construct of evidence (for participants with some training; i.e., novices on the post-test, as well as experts). As a set, these results suggest that context aids the identification of evidence and hypotheses, but may—for situations of low systemic coherence (i.e., considerable controversies)—increase the variability of individuals' ratings of a proposition's believability. Ultimately consistent with this interpretation, pilot studies with experts showed that (a) assessing the present study's context-bound propositions out of (the controversial) context increases the observed reliability of the believability ratings, and (b) employing an in-context situation of high systemic coherence (i.e., of little controversy) yields higher inter-rater reliability for the construct of believability than for no-context propositions.

For no-context propositions, experts showed higher inter-rater reliability for believability, relative to novices. For in-context statements, experts exhibited less (and essentially zero) reliability on their believability ratings, relative to novices. Experts were generally more reliable than novices, as a group, on ratings of hypothesis-likeness. Finally, novices were as reliable as experts on their ratings of evidence-likeness, although this was not initially the case for in-context propositions.

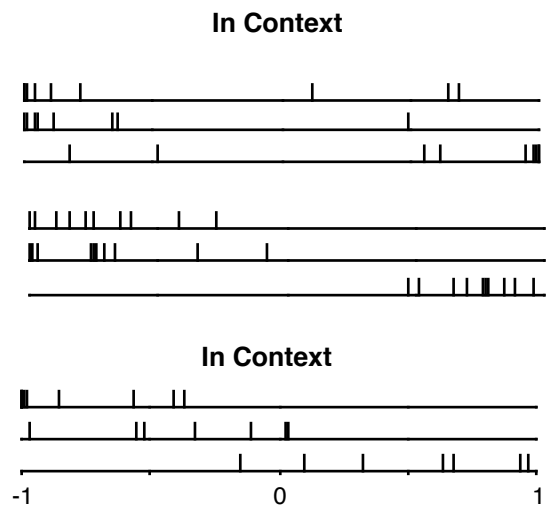
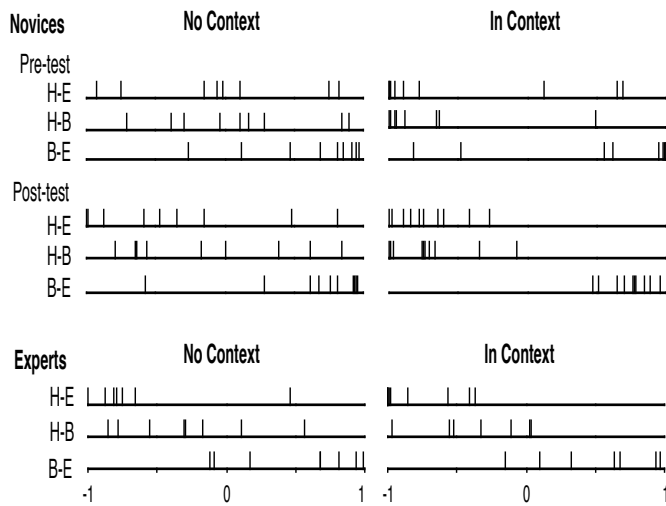
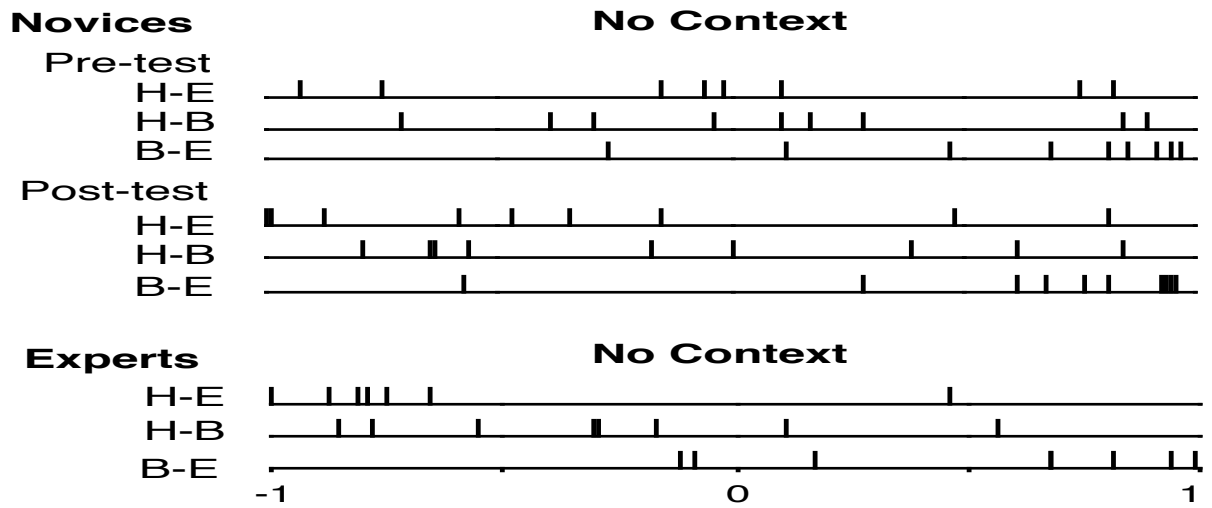


Figure 4.2. Novices' and experts' correlation distributions, Study 1 (from Ranney et al., 1994). H-E, H-B, and B-E refer to hypothesis-evidence, hypothesis-believability, and believability-evidence correlations, respectively.

Individual differences in ratings of the three constructs. Figure 4.2 shows that both experts and novices varied considerably in their approaches to rating the provided propositions. For instance, one novice's initial (no-context) evidence-hypothesis correlation was about $-.9$, while another's was about $.8$; even on the post-test, different novices' correlations (e.g., for believability-hypothesis) yielded ranges of more than 1.5 for three of the six relevant distributions—i.e., the distributions for the no-context propositions. In general, these results mirror many of those mentioned above, as tighter distributions were observed for propositions rated both in context *and* after training; for instance, the participants' believability-evidence correlations ranged only from about $r = .4$ to about $r = 1.0$.

Most instructive, perhaps, are the experts' data. As a group, their correlational distributions had surprisingly wide ranges—sometimes wider than those of the novices. Here again, context seemed to narrow the range of the correlations. As was the case with somewhat fewer novices (especially on their post-test), some experts demonstrated little or no variation for ratings of certain constructs under certain conditions. These were often for principled reasons, even if the principles varied as a function of context and were idiosyncratic with respect to the other participants, including other experts (cf. Ranney, 1994a). For instance, across the no-context propositions, a philosophy professor rated all propositions as the intermediate "5" on (only) the evidence scale, while across the *in-context* propositions, the same expert rated (only) his *believability* consistently as "5." In contrast, another expert rated all in-context propositions as "9" on the believability dimension—that is, as completely believed or accepted.

An interesting case study involves a philosopher of science who rated all no-context propositions as a "1" ("definitely not evidence") on the evidence scale. In a retrospective interview, he explained that the propositions often struck him as "facts" (or sometimes statements of methods), rather than evidence. His distinction (echoed

somewhat by two other experts) was that, by the time something is a *fact*, it is rather theory-independent—while *evidence* counts either for or against a theory. Only one other participant (a novice) showed this data pattern, though, and there certainly seem to be situations in which "facts" are not theory-independent (see the Discussion below).

Other participants were "outliers" in the observed distributions due to near-invariant and/or idiosyncratic responding. For instance, one expert (a professor of cognitive psychology) rated no-context propositions so highly as hypotheses that his evidence- and believability-hypothesis correlations were about 1.1 and .45 higher than those of the next-highest expert, similar to the patterns of some novices. Upon reflection, the expert indicated that, for various imagined scenarios, almost any statement could be viewed as a (perhaps wild or misinformed) hypothesis or "prediction," including the statement, "Abraham Lincoln said that Ross Perot would lose in 1992."

Pilot studies with experts—e.g., involving the testing, at different times, of the in-context propositions without their story context (randomly ordered among other unrelated, isolated propositions)—in conjunction with some of the aforementioned data, indicate that even those that considered or employed principled ways of responding often did so inconsistently (but see Ranney, 1994a and 1994b, for some caveats on metrics of consistency). For instance, in an apparent reversal of his organizing principle, the cognitive psychologist described in the preceding paragraph exhibited rather *negative* correlations between hypothesis-likeness and the other two constructs during the first pilot study—regardless of context. Further, re-testing with a modified corpus of statements appeared to modulate (in this case, considerably reduce) the number of experts who responded in the "principled" fashions described above.

In summary, there appears to be no agreed upon distinction between data and theory; indeed, there is great variability even among experts. Context seems to narrow the decision space, but still leaves many potential roles in which the statement may serve as either a hypothesis or piece of evidence.

Relation Between Novices' Epistemic Categorizations and "Checkbox"

Descriptions

As mentioned in Chapter 3 ("An Example Argument"), when adding (or editing) a statement, novices were also asked to check any number of four descriptions that apply to that statement to help them determine whether the statement was a hypothesis or piece of evidence (see bottom dialog box of Figure 3.2). These descriptions were: (a) "Acknowledged fact or statistic," (b) "Observation or memory," (c) "One possible inference, opinion, or view," or (d) "Some reasonable people might disagree." Statements (a) and (b) were intended to apply more to evidence, and (c) and (d) to apply more to hypothesis, although participants were not told this. They were further asked to choose one of the two (evidence/hypothesis) categories, and to specify the reliability of beliefs classified as evidence. Results indicate that novices' categorizations and checked descriptions were strongly related in the expected direction (see Table 4.3, $p < .001$, omnibus $X^2 = 325.97 > X^2(3) = 16.27$)⁶. Novices selected (a) and (b) more often when categorizing a statement as evidence, and (c) and (d) more often when categorizing it as hypothesis ($p < .001$, $S^* = \sqrt{X^2(3)} = 4.03$). Correlations between the checkbox data (either "1" for checked, or "0" for not checked) and the evidence/hypothesis categorization data (where hypothesis = "0"

⁶One-tailed tests may be justified for these analyses, given that we designed the checkbox descriptions to strongly suggest evidence or hypothesis. However, since some might disagree regarding this issue, two-tailed tests were used.

and evidence = the reliability specified, on a scale from 1-3) are also significant in the expected directions (see last column of Table 4.3). These correlational data suggest that novices view an "acknowledged fact or statistic" as a more reliable piece of evidence than an "observation or memory" ($p < .05$), and view "one possible inference, opinion, or view" as more descriptive of hypotheses than "some reasonable people might disagree" ($p < .05$)—although it's also conceivable that these data reflect an order effect within pairs of checkboxes.

Table 4.3. Frequency and correlational data regarding novices' checked descriptions of a statement, and their categorization of the statement as hypothesis or evidence, Study 1.

<u>Description</u> *	<u>Categorize d as hypo (N=285)</u>	<u>Categorize d as evid (N=251)</u>	<u>Total</u>	<u>Correlation w/ categorization; (H=0, E = 1-3)</u>
(a) Acknowledged fact or statistic	3	112 ^c	115	.56 ^{ab}
(b) Observation or memory	24	124 ^c	148	.44 ^a
(c) One possible inference, opinion, or view	187 ^c	43	230	-.55 ^{ab}
(d) Some reasonable people might disagree	146 ^c	36	182	-.44 ^a
Total	360	315	675	

*For 90 of the 285 hypotheses and 38 of the 251 evidence, no boxes were checked.

^a $r \neq 0$, $p < .001$, 2-tail $Z = 3.28$

^bsignificantly differs from correlation for description (b)/(d) just below, $p < .05$, 2-tail $Z = 1.96$

^csignificantly more selected than for the other (evidence/hypothesis) category, $p < .001$, $S^* = 4.03$

Experts' Ratings of Novices' Definitions

Experts varied considerably in what they considered good definitions of fact, evidence, theory, and hypothesis (as generated by novices), as their respective inter-rater reliabilities were $r = .14, .34, .34,$ and $.39$. (Inter-rater reliabilities for rating definitions of explanation, contradiction, and argument were in a similar range; the agreement on definitions for fact and contradiction did not even differ significantly from zero.) In contrast, inter-rater reliabilities for rating (novices') definitions of less common terms were generally higher (e.g., $.55$ to $.67$ for the notions of confirmation bias, recency bias, and disconfirmation, although the $.21$ agreement on primacy bias was only marginally different from zero). Table 4.4 displays these results, as well as novices' improvements regarding their definitions of the various terms—over half of which are statistically significant. (The novices' mean improvement and mean ultimate performance regarding "recency bias" seem most exceptional.) Ceiling effects for the more common terms may have limited some of these gains.

Table 4.4. Novices' mean pre-test definition scores, post-test change, and intercoder reliability correlations among five (expert) coders, Study 1.

<u>Definition</u>	<u>Pre-test mean score (3 points possible)</u>	<u>Mean changes from pre-test to post-test</u>	<u>Inter-rater reliability (among 5 coders)</u>
hypotheses	2.10	+ 0.13	.39 ^a
evidence	2.11	+ 0.16	.34 ^a
fact	2.04	+ 0.44 ^b	.14
explanation	2.08	- 0.04	.35 ^a
contradiction	2.32	+ 0.10	.12
theory	1.77	+ 0.13	.34 ^a
argument	1.97	+ 0.39 ^b	.37 ^a
confirmation bias	0.88	+ 1.59 ^b	.55 ^a

disconfirmation	0.68	+ 1.50 ^b	.67 ^a
recency bias	0.14	+ 2.49 ^b	.62 ^a
primacy bias	0.00	+ 2.32 ^b	.21 ^c

^a $r \neq 0, p < .05$, 2-tail $Z = 1.96$

^bsignificantly differs from novice's pretest, $p < .05$, 2-tail $T(18) = 2.12$

^c $p = .086$

Exit-Questionnaires and Comments

Mean ratings (on the seven point scale) for what novices thought they "learned the most from" were (in order): *Convince Me* (5.0), Unit 3 (4.3), the integrative exercises (3.89), Unit 2 (3.83), Unit 1 (3.44) and the tests (3.25). These ratings suggest a slight (non-reliable) recency effect, but differences were significant ($p < .05$) only between *Convince Me* and Unit 1, and *Convince Me* and the tests. Table 4.5 presents some comments from the exit questionnaires of the ten novices who used the curriculum with *Convince Me* (three novices did not write any comments about the system).

Table 4.5. Novices' comments about the *Convince Me* system.

Student Comment

- 1 It pays to pause and reflect on your ideas. It should be possible to make it [*Convince Me*] more graphic. That is the program draws a diagram incorporating all the information like one of the diagrams we were asked to draw in one of our exercises.
- 2 It was interesting to see how you could manipulate the data and make it agree better. It doesn't have a visual aid so you can see what you're doing—[add a] graph/chart to turn to?
- 3 Made writing out the argument easier, gave better idea of what sort of models you were trying to evoke.

- 4 I learned how evidence relates to hypothesis. *Convince Me* helped clarify the interrelationship between evidence and hypothesis.
 - 5 I learned a new computer program! Well using the program I did learn that sometimes I seem to be contradicting myself unknowingly. *Convince Me* is pretty neat. I didn't quite understand the algorithm behind it but it's still neat. Could it be made faster?
 - 6 It really helped clarify the argument, the relationships between the pieces and some of the logic. I liked figuring it out, comparing my scores to the computer, I also liked it because sometimes I did well.
 - 7 Really fun.
-

Discussion

Convince Me seems successful at improving novices' abilities to discriminate between the notions of evidence and hypothesis, as suggested by (a) their more negative correlations following training, (b) correlations involving believability that suggest that trained novices take a more positivistic approach, and (c) their comments about the system.

The presence of a context also generally heightened the evidence-hypothesis distinction, although experts seemed less positivistic across context-embedded propositions than the trained novices. Such results are in concert with those of Ranney et al. (1993) and Schank and Ranney (1991), which indicate that context-embedded propositions designed to appear evidence-like are indeed viewed as more believable than propositions designed to appear hypothesis-like. These findings support the Theory of Explanatory Coherence (TEC) and the ECHO model, and seem to conflict with views suggesting that belief evaluation is more likely to proceed top-down, with hypotheses either recruiting evidence or driving theory assessment (e.g., see Mangan & Palmer, 1989). Although some aspects of positivism have fallen into relative disrepute in philosophical circles, people seem to act in accordance with TEC's

principle of data priority, by which—all other things being equal—people are more likely to accept evidence than hypotheses.

That being said, it is perhaps most striking that these results show that it is difficult to determine whether a given statement represents a hypothesis or a piece of evidence. Even for experts, and even for propositions embedded in the context of a story, the inter-rater evidence-likeness agreement was only .42; inter-rater reliability for the hypothesis-likeness construct was only .54. Regarding evidence and hypothesis, several grounds suggest that it is unlikely that people truly "know one when they see one." For one reason, the data presented here indicate that there is great variability even when experts are asked to classify the propositions. For another reason, the task is clearly difficult, often involving much rumination, considerable revision, and conscious reflection before one decides upon ratings for a given statement. Further, the consistency of such ratings does not seem to be high across retesting and changes in context (as is often the case; cf. Ranney, 1994a, 1994b).

To appreciate the difficulty of categorizing even fairly straightforward propositions, consider one of the no-context stimuli, "President John F. Kennedy was assassinated." Many participants saw this as a piece of evidence. For them, the statement is essentially an observation, much like, "This rose is red." In contrast, the philosopher of science saw the proposition as a (nonevidential) fact, a context-free proposition. Yet another interpretation is that it is largely a hypothesis, as the cognitive psychologist maintained. Indeed, there is much to recommend the "hypothesis" perspective. If it were truly a context-free fact, then one could not envision scenarios in which the statement were false. But there are some who may truly believe that (a) the victim was Kennedy's double, (b) Kennedy survived the shooting, or (c) the event was an elaborate suicide. One can envision for nearly any statement a possible situation in which that statement is false or in doubt. So, for such theorists, the statement does appear hypothetical; for one to take it as a domain-

independent fact is to do so *only* for a certain class of theories. Of course, one might suggest that this discussion *merely* turns on some trivial semantic anomaly (e.g., regarding the meaning of "assassinate"). On the contrary, the data collected so far suggest that these considerations are also those of this study's participants. Further, these three divergent responses (i.e., evidence vs. fact vs. hypothesis) were recently independently elicited from each of three Japanese cognitive scientists (who collaborate with one another on related topics)—for the same statement. (Discussions with attorneys about these notions have yielded similar disagreements.) How many of us actually "saw" or "observed" Kennedy's shooting (cf. Hanson, 1958/1965)? It seems fairly clear that people create their own private contexts for such items.

Table 4.6. Some factors that appear to influence a proposition's classification as a hypothesis or piece of evidence (adapted from Ranney et al., 1994).

-
- one's interpretation of the proposition's rhetorical role
 - one's evaluation of the proposition's believability
 - one's interpretation of the proposition's consistency with other beliefs
 - the proposition's grain size of observation
 - one's assessment of the proposition's relative "authority-based" level
 - one's doubt/skepticism
 - one's "epistemology du jour"
 - one's inferences about background context or implicit justifications (i.e., "other" knowledge)
 - one's creativity in recontextualizing or envisioning alternatives
 - one's use of rule-based/logico-deductive reasoning vs. prototypicality-, exemplar- or mental-model-based reasoning
 - one's emotional involvement
 - one's view about what counts as a primitive observation
 - the degree to which one is a reductionist
-

Apparent distinctions between evidence, facts, and hypotheses often appear more clear in the abstract than for concrete cases. "Theory-independent facts" often just means "statements that are part of already-accepted theories" (e.g., "humans are a kind of animal"). What seems like a fact to 20th-century science would likely be a hypothesis—perhaps even heresy—to other people or our own ancestors. Hence, humans-as-animals is only a fact within a class of (sub)theories, so here again context carries the load of what is "indisputable." This helps explain why the rating tasks, particularly for evidence and hypotheses, pose such difficulty. Ranney et al. (1994) proposed a dozen or more factors that influence the categorization of a proposition, but most of these involve aspects of context (see Table 4.6). Further, most of the context must be filled in by the individual, even when the statement is embedded in a story.

The definitional rating data further support the above interpretations, as experts' ratings for novices' definitions of both hypothesis and evidence (as well as "theory" and "explanation") agreed below $r = .4$ (and only .14 for "fact"). It appears that, not only do people not necessarily know a hypothesis (or piece of evidence) when they see it—they may not even be able to agree upon a good definition of it when they see one.

Study 2: Determining the Efficacy of the *Convince Me* Environment

In one sense, *Convince Me* represents a tool, in the same way that a tractor or a word-processor represents a tool. That is, a tractor allows one to plow fields faster and more precisely than one might with a hoe, and a word-processor with a built-in spelling-checker, grammar checker, thesaurus, etc., can enhance the productivity of an author. But it is not obvious that one who has plowed a field with a tractor would be a better plower with a manual plow, or that one who has written a novel using a word-processor would be a better author when returned to paper and pen. However, it's conceivable that the ease of revision afforded by a word processor could help one develop an appreciation for (or habit of) revision that transfers to unsupported practice, and feedback in the form of grammar and spelling checkers could improve these (sub) skills. Similar questions arise regarding *Convince Me*: Does the system's argument development interface and model-driven feedback make its users (a) better reasoners while they employ it, (b) better reasoners even when they are no longer using it, (c) both, or (d) neither?

This second study addresses this question of whether the system is a tool and/or a training device to yield more coherent argumentation skills. Study 2 also sought to determine just how critical *Convince Me*'s knowledge-eliciting interface and simulation-driven feedback are. To do so, two groups were contrasted: a *Convince Me* group that used the software, and a *written* group that received as much of the same instruction as possible, but did their work with paper and pencil (and without feedback—e.g., on the match between their beliefs and ECHO's predictions). The *written* group even received the *Convince Me* manual, but after having read it, was told that the system was "currently unavailable" to them. The results of this study

replicate the essential findings of Study 1 (Ranney et al., 1994) regarding hypotheses and evidence, and as described below, also (a) indicate that the interface and feedback enhanced the students' learning, and (b) demonstrate that the curriculum itself does not account for the full performance gains or positive transfer available via *Convince Me*.

Method

Participants

Twenty novices and four experts participated in this study. The novices (thirteen women and seven men) were undergraduate students from the University of California, Berkeley, who responded to campus advertisements and were paid five dollars per hour for their participation. (The terms "novice" and "student" are again used interchangeably.) As in Study 1 above, the novices' backgrounds were varied, but they had essentially no background in logic or the philosophy of science. The experts were from the University of California (Berkeley): one post-Ph.D. and three doctoral students, two men and two women. Also as in Study 1, the experts had experience in cognitive science, the philosophy of science, science education, and logic, and each is currently studying scientific and practical reasoning. The experts were paid twenty dollars to provide goodness ratings of novices' scientific definitions (which took about two to three hours).

Design and Procedure

The novices completed the pre-test, the three curriculum units on scientific reasoning, the integrative exercises, the post-test, and the exit questionnaire, as described in Chapter 3. However, ten of the novices completed the integrated exercises using the *Convince Me* software (the "*Convince Me* Group," seven women and three men), while the other ten did the exercises with paper and pencil (the

"Written Group;" six women and four men), as shown in Figure 4.3. Both groups were given the same prompts to list/enter hypotheses, evidence, give ratings, etc., and to revise their arguments. The experts were given a randomly-ordered booklet of novices' completed definitions from the pre- and post-tests, and were asked to score each definition on a scale from 1 (poor) to 3 (good).

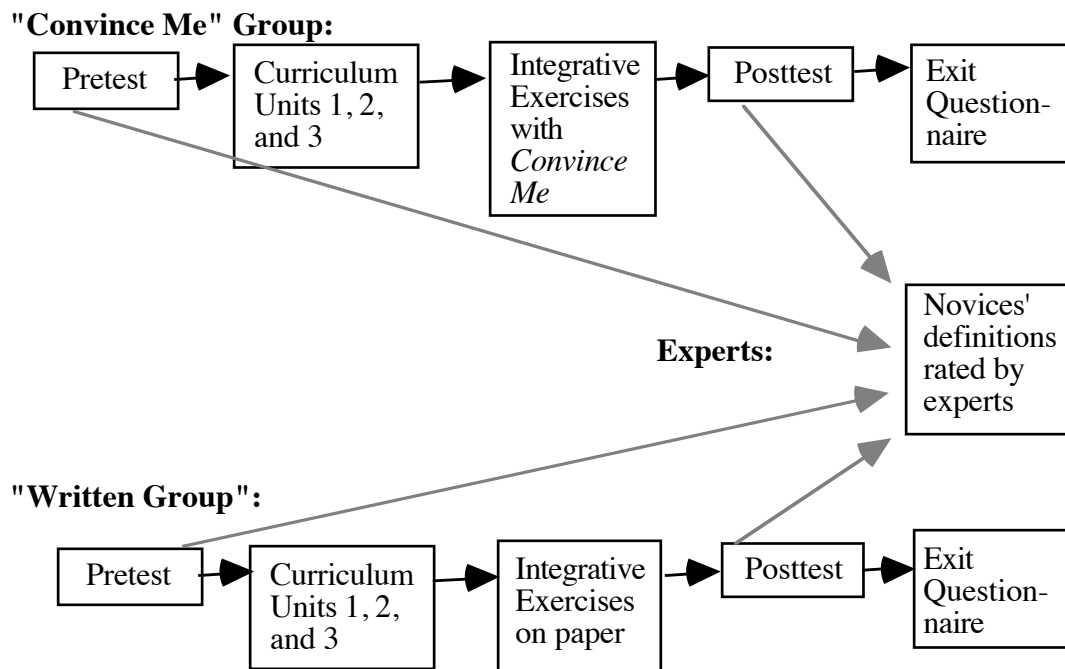


Figure 4.3. Summary of this experiment's method.

Since a portion of this study replicates parts of Study 1, several of its measures were reapplied here. Study 2 also extends Study 1, so additional measures were also employed. For instance, the data were analyzed to ensure that the *Convince Me* and Written groups were not significantly different samples. Further, one striking measure of the utility of *Convince Me* relates to the question of whether students

better articulate and assess their beliefs by virtue of their experiences with the system (vs. using paper and pencil). Put another way, how might one show that individuals' beliefs are more in accord with the structures of their arguments? A relatively straightforward metric comes from the correlation between users' proposition-by-proposition "believability" ratings and the activations generated by an ECHO simulation of the arguments that the individuals generated—whether the ratings (or the arguments) were garnered via *Convince Me* or from paper and pencil. (Of course, this assumes that ECHO is a reasonably accurate model of coherent reasoning. While honest scholars may differ on the extent of that accuracy, Ranney et al., 1993, Schank & Ranney, 1991, 1992, and other work suggest that an increase in the correlation would indeed be indicative of a general improvement in the reflection of one's beliefs in one's arguments.)

Another interesting measure relates to the kinds of changes students employ when making revisions. For instance, do they more often revise their belief ratings, or do they make (perhaps less capricious) changes to the articulation of their argument structure? Compared to the Written group, do *Convince Me* users merely superficially revise their ratings in whatever direction required to provide them with a better correlational fit with ECHO?

Results

Again, two-tailed tests at the $p=.05$ level were conducted for all analyses, unless otherwise noted.

No Differences Between Written and *Convince Me* Samples

Descriptive statistics for the *Convince Me* and Written groups are shown in Table 4.7. As hoped, there were no significant differences between the groups in age,

year in school, SAT scores, or total session hours (two-tailed critical value $t(18) = 2.101$), justifying comparisons between the populations based on the intervention.

Table 4.7. Descriptive statistics for Written and *Convince Me* groups, all measures. Differences between groups were not significant.

<u>Factor</u>	<u>Written</u>	<u>N</u>	<u>Convince Me</u>	<u>N</u>	<u>T-value</u>
Mean age	19.4	10	18.9	10	0.98
Mean year in college	2.3	10	1.6	10	1.64
Mean SAT	1232.0	10	1311.0	9 ^b	-1.84
Math	655.56	9 ^a	697.78	9 ^b	-1.27
Verbal	564.44	9 ^a	613.33	9 ^b	-0.97
Mean hours on task	6.75	10	7.80	10	-1.32

^aOne subject did not report individual Math/Verbal scores

^bOne subject did not report any SAT scores

Propositional Ratings

Correlations among the constructs of evidence, hypothesis, and believability. The results replicate the essential findings of Study 1 regarding hypotheses and evidence for the *Convince Me* users, but not for the Written group (see Table 4.8). The "No Context" data in Table 4.8 include ratings both with and without the isolated "dogs" statements, which were not available for Study 1 (cf. Table 4.1). There were no significant differences between correlations computed with or without the "dogs" statements in the No Context condition, hence comparisons (below) will focus on correlations computed with all statements (i.e., including the isolated "dogs" statements). Unlike in Study 1, context did not seem to significantly affect novices' ability to discriminate between evidence and hypotheses. Note,

however, that the "dogs" statements were seen out-of-context (in question 2) before they were rated in-context (in question 4) in Study 2 (but not in Study 1). Thus, these statements may have been "embedded" in subjects' minds with a certain plausibility before they were given in the story context, possibly diluting any context effect.

Without a context, Written students initially showed a significant correlation between evidence and hypothesis ($r = -.31$), but neither training nor context, nor both together significantly increased the absolute value of the relationship. In contrast, *Convince Me* users also initially show a significant correlation between evidence and hypothesis ($r = -.41$), but training (e.g., $-.63$ and $-.65$), regardless of context, significantly increases the initial relationship. With respect to their believability-hypothesis distinction, the Written students show a nonsignificantly positive correlation ($r = .09$) becoming highly and significantly negative ($-.43$) due to the combined effects of context and training. The *Convince Me* users also show significantly negative believability-hypothesis correlations due to training alone ($-.19$), and context ($-.36$), though these are non-significantly higher than their significantly negative pre-test values ($-.18$ and $-.32$). Believability-evidence correlations did not significantly change from pre- to post-test for either group, but remain generally lower in context compared to no context.

Table 4.8. Within-group correlations between believability and hypothesis-likeness (B-H), evidence-likeness and hypothesis-likeness (E-H), and believability and evidence-likeness (B-E), Study 2.

		Written			Convince Me		
		<u>B-H</u>	<u>E-H</u>	<u>B-E</u>	<u>B-H</u>	<u>E-H</u>	<u>B-E</u>
No Context*	pre	-.09	-.19	.65 ^{ac}	-.17	-.40 ^a	.58 ^{ac}
	post	-.06 ^c	-.08 ^c	.54 ^a	-.11	-.58 ^a	.35 ^a
No context ⁺ :	pre	.09	-.31 ^a	.54 ^{ac}	-.18 ^a	-.41 ^a	.45 ^{ac}
	post	-.14 ^c	-.26 ^a	.47 ^a	-.19 ^a	-.63 ^{ab}	.28 ^a
In context:	pre	-.09	-.38 ^a	.26 ^a	-.32 ^a	-.37 ^a	.05
	post	-.43 ^{ab}	-.39 ^a	.36 ^a	-.36 ^a	-.65 ^{ab}	.12

* Without isolated "dogs" statements (#2 b, d, f, h, j, l, n, p)

+With isolated "dogs" statements (#2 b, d, f, h, j, l, n, p)

^a $r \neq 0, p < .05, 2\text{-tail } Z = 1.96$

^bsignificantly different from pretest, $p < .05, 2\text{-tail } Z = 1.96$

^csignificantly different from in-context (same test), $p < .05, 2\text{-tail } Z = 1.96$

Table 4.9. Between-group correlations regarding believability (B-B), evidence-likeness (E-E), and hypothesis-likeness (H-H), Study 2.

		Written			Convince Me		
		<u>B-B</u>	<u>E-E</u>	<u>H-H</u>	<u>B-B</u>	<u>E-E</u>	<u>H-H</u>
No context*:	pre	.67 ^{ac}	.42 ^{ac}	.11 ^{ac}	.66 ^{acd}	.49 ^a	.27 ^{ac}
	post	.69 ^{ac}	.39 ^{ac}	.02 ^{cd}	.61 ^{acd}	.13 ^{abc}	.10 ^{cd}
No context ⁺ :	pre	.60 ^{ac}	.41 ^{ac}	.22 ^{ac}	.57 ^{ac}	.42 ^a	.34 ^a
	post	.64 ^{ac}	.32 ^{abc}	.18 ^a	.51 ^{ac}	.23 ^{abc}	.36 ^{ac}
In context:	pre	-.01	.19 ^a	.41 ^a	.05	.48 ^a	.42 ^a
	post	-.06	.13 ^a	.25 ^{ab}	.07	.56 ^a	.77 ^{ab}

* Without isolated "dogs" statements (#2 b, d, f, h, j, l, n, p)

+With isolated "dogs" statements (#2 b, d, f, h, j, l, n, p)

^a $r \neq 0, p < .05, 2\text{-tail } Z = 1.96$

^bsignificantly different from pretest, $p < .05, 2\text{-tail } Z = 1.96$

^csignificantly different from in-context (same test), $p < .05$, 2-tail $Z = 1.96$

^dsignificantly different from no-context with "dogs" statements (same test), $p < .05$, 2-tail $Z = 1.96$

Inter-rater agreement regarding the constructs of evidence, hypothesis, and believability. As in Study 1, both groups showed greater inter-rater reliability (correlations) across their believability ratings for the no-context propositions, regardless of testing time (see Table 4.9; cf. Table 4.2). Also as in Study 1, there was generally less agreement regarding the hypothesis-likeness of no-context propositions (relative to in-context propositions) for both groups, and effects in the same direction regarding the construct of evidence for *Convince Me* users with some training (but not for Written students, who displayed less agreement in context). As a set, these results generally replicate those of Study 1 suggesting that context aids the identification of hypotheses but may—for situations of low systemic coherence (i.e., considerable controversies)—increase the variability of novices' ratings of the propositions' believability. Unlike the finding from Study 1, context did not reliably aid the identification of evidence, and even hindered it in some conditions. Further, trained *Convince Me* users generally showed the same or higher inter-rater reliability for evidence-likeness (particularly in-context) and hypothesis-likeness, relative to trained Written students. As a group, Written students were generally as reliable as *Convince Me* users on their ratings of believability.

Regression Analyses

To better understand how perceived evidence- and hypothesis-likeness may influence students' believability ratings, a stepwise multiple linear regression was performed to determine the most parsimonious regression equation for predicting the believability of a proposition. On the pre-test, the full model ANOVA indicates that when hypothesis-likeness, evidence-likeness, and their interaction are included as

predictors in the regression, the resulting equation significantly accounts for 25% of the variability in the believability ratings (see Table 4.10).

Table 4.10. Full Model Regression ANOVA for believability ratings based on the predictors hypothesis-likeness, evidence-likeness, and their interaction.

<u>Source of variation</u>	<u>Degree of freedom</u>	<u>Sum of squares</u>	<u>Mean square</u>	<u>F ratio</u>	<u>R²</u>
Pretest					
Regression	3	551.039	183.680	33.50 ^a	.25
Residual	308	1688.958	5.484		
Posttest: CM					
Regression	3	76.063	25.354	4.26 ^a	.08
Residual	150	892.879	5.953		
Posttest: Written					
Regression	3	242.272	80.757	14.33 ^a	.22
Residual	152	856.875	5.637		

^asignificant at $p < .05$, Reject H_0 : $B_{Hyp} = B_{Evid} = B_{Evid*Hyp} = 0$, if $F > F_{2,308;.95} = 3.00$

A post-hoc analysis of the significance of the slopes for each predictor reveals that two of the predictors—evidence-likeness and the constant—contribute significantly in accounting for the total variance in believability ratings, while the interaction between evidence-likeness and hypothesis-likeness makes a marginal contribution. The effect of perceived hypothesis-likeness on students' believability ratings was not found to be significant ($p = .11$; see Table 4.11). On the post-test, evidence-likeness and the constant were again the only significant predictors of

believability ratings. Compared to the pre-test, however, the best regression equations accounted for less of the variability in believability ratings (22% for Written subjects, 8% for *Convince Me* subjects; see Tables 4.10 and 4.11). Thus, it appears that believability is more so correlated with evidence-likeness than anti-correlated with hypothesis-likeness (supporting notions of data-priority and positivism). Further, the structure of the argument (rather than just the epistemic categorizations) may account for a fair amount of the remaining 75% or so of the variance in believability ratings.

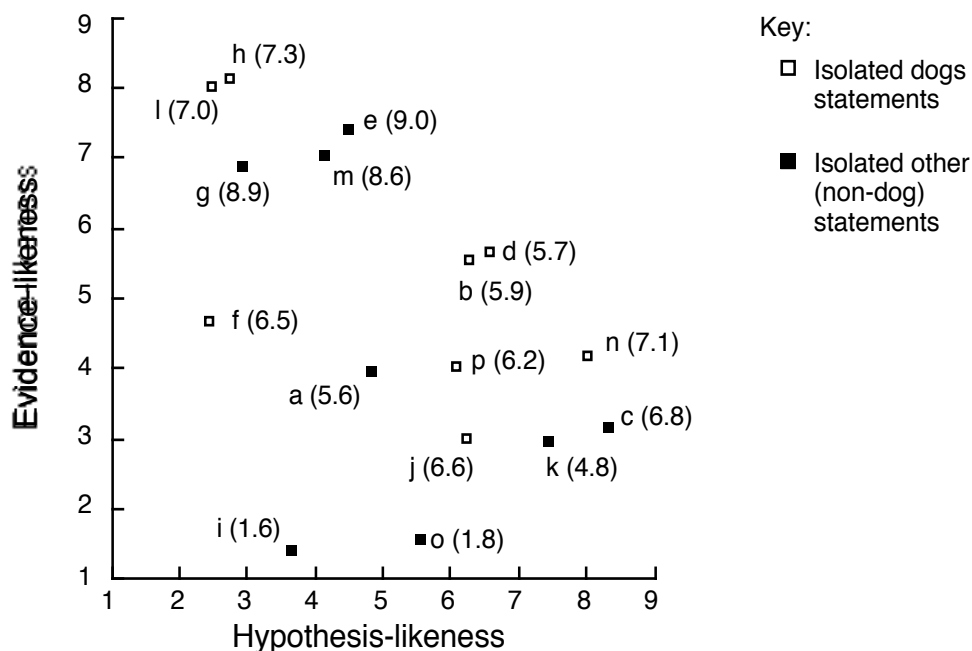
Table 4.11. Post-hoc tests of regression slopes for each potential believability rating predictor—hypothesis-likeness (H), evidence-likeness (E), their interaction (H*E), and a constant.

<u>Predictor Variable</u>	<u>Slope B</u>	<u>SE B</u>	<u>T Value</u>	<u>p (2 Tail)</u>
Pretest				
1. H	0.112	0.070	1.593	0.112
2. E	0.512	0.072	7.165 ^a	0.000
3. H * E	-0.021	0.012	-1.783	0.076
4. Constant	3.586	0.477	7.511 ^a	0.000
Posttest—CM				
1. H	0.083	0.119	0.697	0.487
2. E	0.302	0.123	2.449 ^a	0.015
3. H * E	-0.025	0.021	-1.214	0.226
4. Constant	4.740	.923	5.138 ^a	0.000
Posttest—Written				
1. H	0.063	0.139	0.453	0.651
2. E	0.492	0.139	3.534 ^a	0.001
3. H * E	-0.014	0.020	-0.661	0.510
4. Constant	3.657	1.006	3.637 ^a	0.000

^a significant at $p < .05$

Two-Dimensional Evidence- and Hypothesis-likeness Plot

Figure 4.4 is a plot of mean evidence- versus hypothesis-likeness for the isolated propositions in the pre-test (for all novices, i.e. combining Written and *Convince Me*), labeled by statement and mean believability rating. As shown, the statements cover most of the space: Some were rated high on hypothesis-likeness and low on evidence-likeness (e.g., c, k), some were the reverse (high on evidence-likeness and low on hypothesis-likeness, e.g., h, l), some were rated comparably on both dimensions (e.g., a, b, d), and some were rated low on both dimensions (e.g., i). In general, the non-dogs statements varied a lot in terms of believability, with means ranging from 1.6 to 9.0, correlating mostly with evidence-likeness (consistent with the data in Table 4.8). The believability of the dogs statements varied much less (and seemed to contribute more "noise"), with means ranging from 5.7 to 7.3.



<u>Statement</u>	<u>H</u>	<u>E</u>	<u>B</u>
a) All wine is made from grapes.	4.9	4.0	5.6
b) Some dogs have an aggressive disorder in which they bark more, growl more, bite more, and have higher blood pressure and heart rate than other dogs do.	6.3	5.6	5.9
c) Gravity exists in other galaxies.	7.5	3.0	6.8
d) Lack of a chemical causes an aggressive disorder in dogs.	6.6	5.7	5.7
e) Gravity exists on Earth.	4.5	7.4	9.0
f) Some researchers trained one group of aggressive-disorder dog owners to treat their dogs firmly yet lovingly.	2.5	4.7	6.5
g) President John F. Kennedy was assassinated.	3.0	6.9	8.9
h) Some researchers found that training dog owners to treat their dogs firmly yet lovingly relieved symptoms of aggressive disorder in their dogs.	2.8	8.1	7.3
i) Abraham Lincoln said that Ross Perot would lose in 1992.	3.7	1.4	1.6
j) Some researchers think dogs get an aggressive disorder when their owners treat them poorly.	6.3	3.0	6.6
k) Birds evolved from animals that lived in trees.	8.4	3.2	4.8
l) Other researchers found that a chemical relieved symptoms of aggressive disorder in dogs.	2.5	8.0	6.9
m) Approximately three-quarters of the surface of the Earth is covered by water.	4.2	7.0	8.6
n) Abuse causes an aggressive disorder in dogs.	8.1	4.2	7.1
o) All humans on Earth are dead at this moment.	5.6	1.6	1.8
p) Other researchers think that dogs get an aggressive disorder because they lack a certain chemical.	6.1	4.0	6.2

Figure 4.4. Plot (and listing) of propositions' mean ratings along the hypothesis-likeness and evidence-likeness axes, with mean believability (in parentheses), Study 2.

Table 4.12. Frequency and correlational data regarding novices' checked descriptions of a statement, and their categorizations of a statement as hypothesis or evidence, Study 2.

<u>Description</u> *	<u>Categorized as hypo (N=237)</u>	<u>Categorized as evid (N=274)</u>	<u>Total</u>	<u>Correlation w/ categorization; (H=0, E = 1-3)</u>
(a) Acknowledged fact or statistic	7	157 ^c	164	.69 ^{ab}
(b) Observation or memory	41	161 ^c	202	.41 ^a
(c) One possible inference, opinion, or view	204 ^c	70	274	-.68 ^{ab}
(d) Some reasonable people might disagree	138 ^c	43	181	-.52 ^a
Total	390	431	821	

*For 17 of the 237 hypotheses and 11 of the 274 evidence, no boxes were checked.

^a $r \neq 0$, $p < .001$, 2-tail $Z = 3.28$

^bsignificantly different from correlation for description (b)/(d) just below, $p < .001$, 2-tail $Z = 3.28$

^csignificantly more selected than for the other (evidence/hypothesis) category, $p < .001$, $S^* = 4.03$

Relation Between Epistemic Categorizations and "Checkbox" Descriptions

As in Study 1, when adding (or editing) a statement, *Convince Me* users were also asked to check any number of four descriptions to help them determine whether the statement was a hypothesis or a piece of evidence, to select one of the two (evidence/hypothesis) categories, and to specify the reliability of beliefs classified as evidence (see Chapter 3, "An Example Argument"). Novices' categorizations and checked descriptions were strongly related as expected, and mirrored Study 1's findings, as shown in Table 4.12 (cf. Table 4.3; $p < .001$, omnibus $X^2 = 330.67 > X^2(3) = 16.27$)⁷; novices selected (a) or (b) more often when categorizing a statement as evidence, and (c) and (d) more often when categorizing it as hypothesis ($p < .001$, $S^* = \sqrt{X^2(3)} = 4.03$). As in Study 1, correlations between the checkbox data (where checked = "1" and not checked = "0") and the evidence/hypothesis categorization data (where hypothesis = "0" and evidence = the specified 1-3 reliability) are also significant in the expected directions. The correlational data indicate (again) that novices see (a) an "acknowledged fact or statistic" as a more reliable piece of evidence than (b) an "observation or memory" ($p < .001$), and (c) "one possible inference, opinion, or view" as more descriptive of hypotheses than (d) "some reasonable people might disagree" ($p < .001$; see Table 4.12).

Experts' Ratings of Novices' Definitions by Group

As in Study 1, experts had considerable variety in what they considered good definitions (by novices) of the common terms (e.g., fact, evidence, theory, and hypothesis, etc.; cf. Table 4.4). Table 4.13 displays these results for both groups, as well as novices' improvements regarding their definitions of the various terms. Over

⁷Again, one-tailed tests may be justified for these analyses, given both our strong expectations and prior study results—but since some may disagree regarding this issue, two-tailed tests were used.

half (six of eleven, as with Study 1) of these pre- to post-test gains are statistically significant for *Convince Me* users. For Written students, four of the eleven gains were significant. Again, ceiling effects for the more common terms may have stunted some of these gains. *Convince Me* users' gains were numerically larger than Written students' gains for seven of the eleven definitions. Inter-rater reliabilities were also similar to Study 1's, with a surprisingly low overall reliability of $r = .48$ ($p < .05$)—again suggesting that there is great variability even among experts regarding the meanings of common reasoning terms. As in Study 1, inter-rater reliabilities for less common terms were generally higher than for more common terms, and ranged between $r = .18$ for "fact," and $r = .70$ for "primacy bias."

Table 4.13. Novices' mean pre-test definition scores, post-test change, and intercoder reliability correlations among four (expert) coders, Study 2.

<u>Definition</u>	<u>—Convince Me—</u>		<u>—Written—</u>		<u>Overall inter-rater reliability</u>
	<u>Pre-test mean</u>	<u>Mean changes pre- to post-test</u>	<u>Pre-test mean</u>	<u>Mean changes, pre- to post-test</u>	
hypotheses	2.08	+0.36 ^b	2.41	+0.05	.37 ^a
evidence	2.01	+0.38 ^b	2.29	+0.22	.49 ^a
fact	2.35	+0.02	2.55	-0.07	.18 ^a
explanation	1.99	+0.12	1.84	+0.31	.22 ^a
contradiction	2.36	-0.02	2.17	+0.11	.14 ^a
theory	1.69	+0.20	1.68	+0.35 ^b	.33 ^a
argument	1.85	+0.17	1.94	0	.52 ^a
confirmation bias	1.03	+1.58 ^b	1.41	+0.83 ^b	.63 ^a
disconfirmation	1.04	+1.24 ^b	1.43	+0.22	.50 ^a
recency bias	1.06	+1.15 ^b	0.49	+1.64 ^b	.67 ^a
primacy bias	0.93	+1.52 ^b	0.52	+1.46 ^b	.70 ^a

^a $r \neq 0, p < .05, 2\text{-tail } Z = 1.96$

^bsignificantly differs from novice's pretest, $p < .05, 2\text{-tail } T(18) = 2.12$

Argument Revisions

After completing an initial argument (and running a simulation, for the *Convince Me* subjects), subjects were asked to make any revisions to their initial arguments or ratings that seemed appropriate. An argument revision was defined as an episode of adding, editing, or deleting any number of hypotheses, evidence, explanations, or contradictions in an initial argument. Similarly, a rating revision was defined as an episode of changing one or more ratings. Given these definitions, Table 4.14 shows that during the main exercises, students in both groups made about the same *total* number of changes to their arguments. However, on the exercises and overall, *Convince Me* users modified their argument structures twice as often as their ratings—while Written students did the *reverse*, changing their ratings twice as often as their arguments ($p < .05$, omnibus $X^2(1) = 3.84$, $S^* = \sqrt{X^2(1)} = 1.96$; see Figure 4.5). (On the pre-test and post-test, both groups again made about the same small number of modifications, and both made nonsignificantly more changes to ratings than to arguments).

Table 4.14. Changes to arguments and ratings, *Convince Me* and Written groups.

<u>Change</u>	<u>Pretest</u>	<u>Exercises</u>	<u>Posttest</u>	<u>Overall</u>
<i>Written</i>				
Arguments	1	8	0	9
Ratings	2	13 ^a	3	18 ^a
TOTAL	3	21	3	27

*Convince Me**

Arguments	1	16	1	18
Ratings	1	7 ^a	2	10 ^a
TOTAL	2	23	3	28

*Only one user changed parameter settings during the exercises (9 changes).

^asignificantly different number of changes to ratings as to arguments, within group, $p < .05$, $S^* = 1.96$

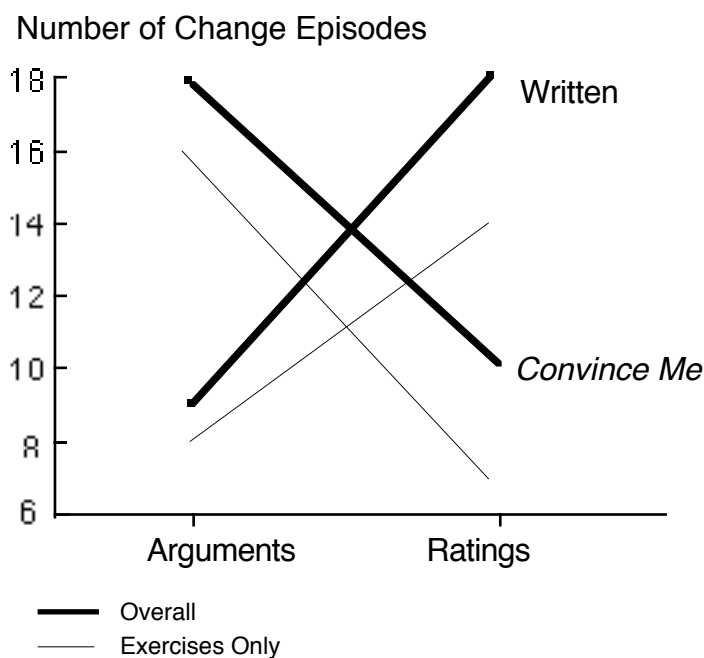


Figure 4.5. Argument and rating change episodes, *Convince Me* and *Written* groups.

Thus, *Convince Me* users apparently don't just try to "mimic" ECHO by changing their ratings. On the contrary, compared to students developing arguments on paper, students using the system seem more likely to reflect on and change the fundamental structure of their arguments. *Convince Me* seems to offer useful support for structuring and revising arguments, beyond that offered by paper and pencil. This may be somewhat due to *Convince Me* subjects both (a) getting feedback, and (b) knowing they'd be getting feedback (i.e., during the exercises).

Believability-Activation Correlations

Table 4.15 shows the overall belief-activation correlations for novices' arguments. Using paper-and-pencil pre-test items, the belief-activation correlations aggregated over both the Written and the *Convince Me* groups were around (and numerically under) .3. However, during the main exercises (which follow the curriculum), students who used the software significantly improved their correlation to .62. Even the correlation for initial arguments, before users were given a chance to make revisions based on ECHO's feedback, were significantly improved from the pre-test—also to about .6. The Written group's instruction also reduced the competence/performance gap, significantly improving their correlation to .47—although this value is significantly lower than the .62 evidenced by the *Convince Me* group.

Table 4.15. Overall belief-activation correlations on the first argument, the last revised argument, and all arguments.

	<u>Pretest r</u>	<u>Exercises r</u>	<u>Posttest r</u>	<u>Overall r</u>
<u>First argument</u>				
<i>Written</i>	.30 ^a	.45 ^{ab}	.39 ^a	.41 ^a
<i>Convince Me</i>	.30 ^a	.61 ^{abd}	.49 ^{acb}	.51 ^{ad}
<u>Last argument</u>				
<i>Written</i>	.33 ^a	.50 ^{ab}	.39 ^a	.43 ^a
<i>Convince Me</i>	.29 ^a	.68 ^{abd}	.49 ^{abc}	.56 ^{ad}
<u>All arguments</u>				
<i>Written</i>	.31 ^a	.47 ^{ab}	.38 ^a	.42 ^a
<i>Convince Me</i>	.24 ^a	.62 ^{abd}	.51 ^{ab}	.53 ^{ad}

^ar ≠ 0, p < .05, 2-tail Z = 1.96

^bsignificantly higher than pretest, within group, $p < .05$, 1-tail $Z = 1.64$

^csignificantly lower than exercises, within group, $p < .05$, 1-tail $Z = 1.64$

^dsignificantly different from other group, same activity, $p < .05$, 2-tail $Z = 1.96$

But do these instructional methods also yield transfer? For the Written group, at least over the period of this short intervention, the answer is certainly "not much;" the belief-activation correlation for their post-test dipped to .38, which was nonsignificantly higher than their pre-test performance. In marked contrast, the *Convince Me* group maintained a significant post-test advantage (with a correlation of .51) over their own pre-test correlation, which did not significantly dip during the post-test (i.e., when they no longer had access to the software), and maintained a marginal advantage over the Written group's post-test correlation ($z=1.75$, two-tailed $p=.08$; see Figure 4.6.)

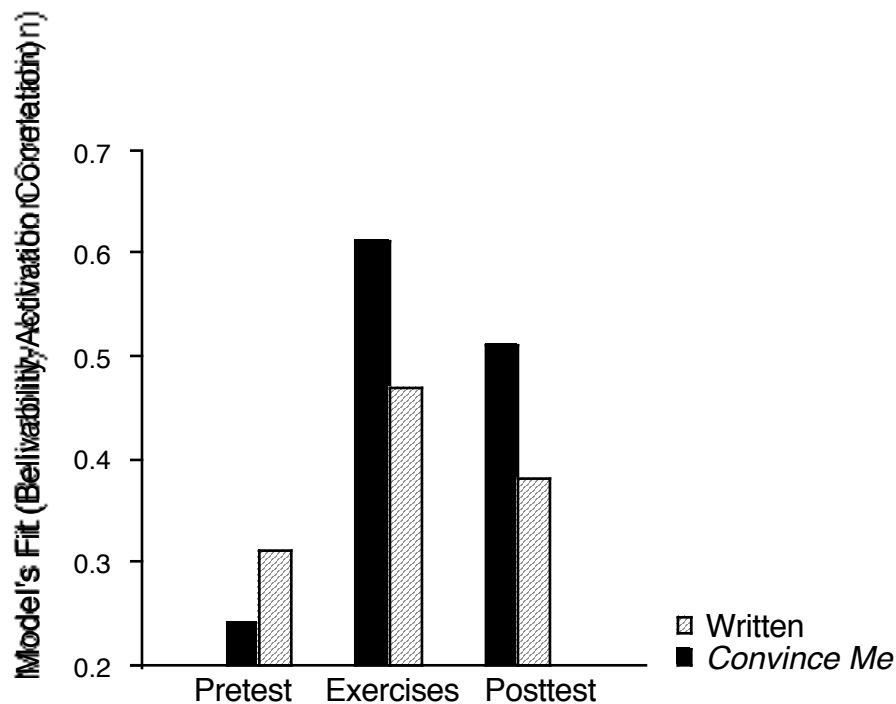


Figure 4.6. Overall model's fit, all arguments, *Convince Me* and Written groups (from Schank & Ranney, 1995).

Argument Analyses

How do *Convince Me* and Written students' arguments differ? For instance, do they elaborate them in different ways (e.g., do they focus on explicating their own argument, versus generating competing hypotheses)? When students generate their own propositions, how proficient are they at categorizing them as hypothesis or evidence (i.e., are their categorizations questionable)?

Table 4.16. Believability-activation correlations and the mean number of hypotheses, evidence, new propositions (hypotheses or evidence that were not in the given text/situation), explanations (including joint explanations, in parentheses), and contradictions for novices' arguments—overall (across tests and integrative exercises), and for the tests and integrative exercises separately.

<u>Argument</u>	<u>r</u>	<u>Hypo-</u> <u>theses</u>	<u>Evi-</u> <u>dence</u>	<u>New</u> <u>propo-</u> <u>sitions</u>	<u>Explanations</u> <u>(and joint</u> <u>explanations)</u>	<u>Contra-</u> <u>dictions</u>
Written						
OVERALL	.42	5.4	5.9	3.8	8.6 (0.2)	4.7
Pretest:						
Wine	.50	5.4	5.2	2.2	5.6 (0.0)	1.2
Language	.52	5.4	4.6	3.0	4.0 (0.0)	1.4
HIV	.34	4.8	6.0	2.4	5.2 (0.0)	3.2
Movie	.32	7.6	7.6	4.2	10.0 (0.0)	8.4
Exercises:						
Yawn	.58	6.0	8.3	3.2	10.3 (0.9)	2.3
Glumpis	.57	6.5	6.1	4.3	9.0 (0.1)	6.3
Pendulum	.80	4.2	4.1	7.0	8.3 (0.6)	5.4
Abortion	.34	4.5	4.8	4.7	6.6 (0.1)	4.9
Posttest:						
Wine	.28	4.0	5.4	2.4	6.6 (0.0)	4.6
Language	.43	5.2	5.6	2.6	10.4 (0.0)	4.2
HIV	.53	5.2	6.8	2.6	12.4 (0.0)	8.8
Movie	.35	6.2	7.0	3.4	15.4 (0.0)	5.0
<i>Convince Me</i>						
OVERALL	.53 ^a	3.9 ^a	5.5	2.4 ^a	10.6 (1.4)	5.8
Pretest:						
Wine	.68	2.5 ^a	4.3	1.3	3.8 (0.0)	2.0

Language	.55	3.8	4.3	2.3	5.7 (0.0)	2.2
HIV	.31	3.2	5.3	1.5	5.5 (0.0)	5.2
Movie	.46	4.0 ^a	6.8	1.8 ^a	12.0 (0.0)	3.5
Exercises:						
Yawn	.70	4.0 ^a	7.8	0.5 ^a	17.8 (3.8)	7.0 ^a
Glumpis	.73	4.0 ^a	5.3	0.6 ^a	15.9 (3.1)	7.0
Pendulum	.88	4.1	3.9	7.3	9.9 (2.4)	10.6
Abortion	.81 ^a	3.5	4.3	3.6	10.2 (1.6)	6.0
Posttest:						
Wine	.68 ^a	3.0	6.0	1.2	8.3 (0.3)	3.0
Language	.49	4.8	5.3	2.0	7.3 (0.0)	2.3
HIV	.69	4.5	5.8	0.3 ^a	7.0 (0.0)	2.5
Movie	.42	6.0	7.3	3.3	11.7 (0.0)	9.5

^asignificantly different from Written, same condition, $p < .05$, 2-tail $t = 2.12$

Descriptive statistics for novices' arguments are given in Table 4.16. Between groups, arguments did not differ in terms of the quantity of evidence used, but they did differ in the following ways: First, *Convince Me* users showed numerically higher believability-activation correlations overall on all exercises and post-test arguments, significantly so for the abortion exercise (.81 vs. .34) and wine post-test argument (.68 vs. .28). Second, Written students used significantly more hypotheses in half of the pre-test arguments, and in the first two exercises. Third, Written students introduced slightly more new propositions (mainly hypotheses) in general (i.e., statements unmentioned in the given text/situation) than *Convince Me* users. Fourth, *Convince Me* users tended to use more explanations and contradictions in their arguments overall, significantly so when using the software (21.1 vs. 13.3, $p < .05$), and on the "yawn" exercise (24.8 vs. 12.6, $p < .05$). *Convince Me* subjects had also marginally more explanations overall (10.6 vs. 8.6, $p = .066$), and significantly more contradictions on the "yawn" exercise (7.0 vs. 2.3, $p < .05$).

These results suggest that without *Convince Me* users may generate slightly more alternate, competing hypotheses, but *Convince Me* helps students better clarify and connect their argument (e.g., make the explanatory and contradictory relations more explicit)—while still supporting their generation of competing hypotheses. This partly explains why *Convince Me* users' belief-activations correlations were higher—more explicated arguments should better reflect underlying beliefs.

The abortion argument was more closely examined since it was the last argument novices completed in the integrative exercises (before the post-test), and the textual stimulus provided was minimal (two sentences, with about four or fewer propositions). Mainly due to the detail and time required for these analyses, they was performed by the author alone (i.e., inter-rater data was not collected).

On the abortion argument, five of the twenty Written and *Convince Me* students had pro-life leanings, ten had pro-choice leanings, and five were fairly

neutral on the issue (with two of these slightly leaning pro-life, and one slightly leaning pro-choice; see Table 4.17, fourth column). These leanings were uncorrelated with group membership. Five of the twenty arguments—all for Written students—had fairly low or negative believability-activation correlations, both absolutely and relative to their own overall correlations without the abortion case; of these arguments, one was pro-life, two were neutral/slightly pro-life, and two were pro-choice (see students DE, FE, HA, LI, and LA in Table 4.17). Thus, low believability-activation correlations were fairly evenly distributed between pro-life, pro-choice, or neutral arguments—although none were observed in the *Convince Me* group. Unlike the Written subjects, all of the *Convince Me* subjects' abortion argument correlations were as high or higher than their overall correlations without the abortion case. Further, students generated numerically (but non-significantly) more propositions for "their side" of the argument (the ratio of "my side" versus "other side" propositions was 1.13), and there were no significant differences between groups on this measure (see Table 4.18).

Table 4.17. Believability-activation correlations for all novices, overall without the abortion case, and on the "abortion" argument (for their last revision), believability ratings for the "abortion is okay" proposition versus the "abortion is wrong" proposition, number of pro ("okay") versus con ("wrong") abortion propositions, and total number of argument propositions and links.

<u>Student</u>	<u>Overall r's without abortion_ case</u>	<u>Abortion r's</u>	<u>Abortion as "Okay" vs. "Wrong" Ratings</u>	<u>Number of "Okay" vs. "Wrong" Props</u>	<u>Total Props and Links</u>
<i>Written</i>					
DE	.47 ^a	-.59	2 : na	4 : 2	6,4
WA	.54 ^a	.78	1 : 9	5 : 3	8,10
JO	.27 ^a	.56	9 : 3	8 : 4	12,22
FE	.44 ^a	.10	5 : 6	6 : 2	8,7
DI	.36 ^a	.83 ^a	2 : 8	4 : 4	8,7
ST	.51 ^a	.69	6 : 4	5 : 2	7,7
HA	.50 ^a	.14	6 : 7	5 : 4	9,9
LI	.59 ^a	.29	8 : 1	4 : 5	9,15
LA	.29 ^a	-.23	7 : 3	10 : 5	15,27
IM	.56 ^a	.86 ^a	8 : 2	4 : 6	10,7
<i>Convince Me</i>					
RI	.49 ^a	.78 ^a	6 : 2	5 : 4	9,11
MA	.59 ^a	.91	2 : 8	2 : 2	4,5
VE	.22 ^a	.92 ^{ab}	7.5 : 2.5	5 : 5	10,11
EL	.64 ^a	.84 ^a	8.4 : 1.5	6 : 3	9,9
SI	.66 ^a	.85 ^a	7 : 3	7 : 5	12,48
ME	.78 ^a	.97 ^{ab}	2 : 8	3 : 5	8,10
SA	.58 ^a	.73	7 : 3	4 : 2	6,9
SL	.51 ^a	.61	7 : 7	2 : 2	4,6
CH	.62 ^a	.56 ^a	5 : 5	5 : 5	10,40
KE	.22 ^a	.88 ^{ab}	4 : 1	3 : 3	6,12

^a $r \neq 0, p < .05, 2\text{-tail } Z = 1.96$

^bsignificantly different from overall r without abortion case, $p < .05, 2\text{-tail } Z = 1.96$

Table 4.18. Number of "my side" versus "other side" propositions for the abortion argument, by group. (Differences are not significant, $X^2 = .70$.)

<u>Group</u>	<u>My Side</u>	<u>Other Side</u>	<u>Total</u>
Written	46	46	92
<i>Convince Me</i>	44	34	78
Total	90	80	170

Ratio of "my side" versus "other side" arguments: 1.13

In general, the author (in a blind categorization) agreed with students' categorizations of their self-generated propositions (as hypothesis or evidence) more than might be expected. Overall, she questioned the categorization of about 16% of the students' statements—about 8% of novices' hypothetical classifications, and about 23% of their evidential ones (see Table 4.19, bottom rows). The questionability of categorizations did not differ significantly between the *Convince Me* and Written groups. Disagreement was not distributed evenly across students—22% of the novices accounted for 52% of the questionable categorizations, while for 1/3 of the students, *none* of their categorizations appeared questionable.

The most common difficulty occurred when students categorized an assertion as a piece of evidence when the author viewed it as hypothesis. Across all students on the abortion argument, 21 statements were questionable along this line, as shown in Table 4.19. For instance, one student classified the ("pro-choice") statements "Population is too high," "Unwanted children only cause more problems," and "There are many who would adopt unwanted children" as evidence, when all three seem clearly arguable (partly because they include vague quantifiers). Another student listed as ("anti-abortion") evidence "It is a personal not societal issue," "Everyone has a God-given right to live," "Right of the fetus [to live]," and "Babies

with terminal diseases better off aborted" which, again, are some of the more controversial (and arguable) assumptions at the center of the debate! Examples from other students include: "Our society always follows the guidelines of the Bible," "Fetuses are alive, but they have no consciousness," "Forcing women to have children they do not want is bad," and "Killing is not so bad." The checkbox data for these statements (see "Relation between epistemic categorizations and "checkbox" descriptions," above) support their hypothesis-likeness: (c) "One possible inference, opinion, or view" and (d) "Some reasonable people might disagree" (suggesting hypothesis) together were checked a total of 13 times, while (a) "Acknowledged fact or statistic" and (b) "Observation or memory" (suggesting evidence) together were checked only 5 times (see Table 4.19).

Table 4.19. Questionability of evidence/hypothesis categorization for the "abortion" argument.

	<u>No. of Props</u>	<u>No. Questionable</u>	<u>% Questionable</u>
<u><i>Convince Me</i></u>			
Hypotheses	35	2	5.7%
checkboxes $\sqrt{}$ ed:		a-0 b-2 c-2 d-1	
Evidence	43	10	23.2%
checkboxes $\sqrt{}$ ed:		a-3 b-2 c-8 d-5	
<u>Written</u>			
Hypotheses	45	4	8.9%
Evidence	48	11	22.9%
<u>TOTAL</u>			
Hypotheses	80	6	7.5%
Evidence	91	21	23.1%
Props	171	27	15.8%

Less common were questionable categorizations of hypotheses that seemed more like evidence, such as "Fetuses are alive" (they are made of living tissue—the question is whether they are "people" with certain "rights") and "We as a society kill living things" (we clearly kill animals for food, we kill some criminals, etc.). These two statements accounted for all but one of the (six total) "questionable" hypotheses across all novices for this argument. (The sixth was "Children raised by people who do not care for them can suffer physically and/or emotionally"—a fairly well-accepted observation; the more contestable, related issues seem to be whether such children should be taken from their parents via foster care, adoption, etc., or whether the right to life is worth taking the chance of growing up in an impoverished environment.) The checkbox data for these statements fairly equally support either categorization, as checkbox (a) and (b) (suggesting evidence) together were checked a total of 2 times, while (c) and (d) (suggesting hypothesis) together were checked a total of 3 times (see Table 4.19).

Exit-Questionnaires and Comments

Mean ratings for how much students thought they learned from the various activities are given in Table 4.20. For *Convince Me* users, mean ratings were (in order): *Convince Me* (5.7), Unit 3 (5.0), Unit 1, Unit 2, and the tests (tied at 4.5), and the integrative exercises (4.3). These differences were significant ($p < .05$) only between *Convince Me* and the exercises. These results were similar to Study 1 in that only *Convince Me* was rated significantly higher than at least one other activity—although this could be partially attributed to a "halo" effect of technology. For Written students, the means (in order) were: Unit 2 (5.8), Unit 1 (4.8), the exercises and tests (tied at 4.7), Unit 3 (4.5), and *Convince Me* (4.17). (Only 6 of the 10 Written students offered ratings for *Convince Me*, and with high variation, presumably since they didn't use the system.) These differences were significant

($p < .05$) only between Unit 2 and Unit 1, and Unit 2 and the exercises. Table 4.21 and 4.22 present students' comments from the exit questionnaires for the *Convince Me* and Written groups, respectively. (One of the Written students did not offer comments.) Together, the ratings and comments suggest that when students have the opportunity to use *Convince Me*, they view it (along with its Unit 3 manual) as the most useful activity; but when the system is not available, they (i.e., the Written group) view Unit 2 as the most useful, although they see *Convince Me* as a tool they would *like* to use.

Table 4.20. How much novices thought they learned (on a 1-to-7 scale, in which 1 = not much and 7 = a lot).

<u>Activity</u>	<u>Mean</u>	<u>N</u>	<u>sd</u>	<u>Max</u>	<u>Min</u>
<i>Convince Me</i>					
Unit 1	4.5	10	1.18	7	3
Unit 2	4.5	10	1.27	6	3
Unit 3	5.0	10	1.25	7	3
<i>Convince Me</i>	5.7 ^a	10	1.42	7	3
Exercises	4.3	10	1.41	7	2
Tests	4.5	10	1.78	7	1
<i>Written</i>					
Unit 1	4.8	10	1.03	6	3
Unit 2	5.8 ^b	10	0.79	7	5
Unit 3	4.5	10	1.84	7	2
<i>Convince Me</i>	4.17	6	1.34	7	1
Exercises	4.7	10	1.34	7	3
Tests	4.7	10	2.00	7	1

^asignificantly different from Exercises (same group), $p < .05$, 2-tail $T(18) = 2.101$

^bsignificantly different from Unit 1 and the Exercises (same group), $p < .05$, 2-tail $T(18) = 2.101$

Table 4.21. *Convince Me* users' comments about the system, after using it.

Student Comment

- 1 It's fun. you can change your arguments and evidence to make the computer see your point of view...The computer helped me create the arguments, I liked having the structure, it made it easier to make the connections...*I always thought it was pretty clear what a hypothesis or piece of evidence was, that it was a pretty formal thing. I was surprised how fuzzy they actually are...If you'd asked me before the test I would have thought distinguishing hypotheses from evidence would have been an easy task, I was surprised that it wasn't....But it got easier after using the program.*
- 2 I enjoyed it, especially doing the tests and using *Convince Me*. I think I did a lot better on the post-test, *things were a lot clearer in my mind, I liked reflecting on my thinking...I think I did better on the last test mostly because of using the program.*
- 3 I learned a lot about what's needed to make a good argument--it's pretty tough!
- 4 I liked to see how the argument was interpreted. I learned how to formulate an argument, *how to organize evidence and hypotheses to tell me something.*
- 5 The program made clearer for me the way I think, also made clearer for me what a strong logical argument needs. I learned that some things I believe strongly in, I cannot argue well, and I have a confirmation bias.
- 6 It's interesting to find out how my ratings compared to the computer's.
- 7 Really illustrates logic processes and makes you see the holes in arguments. *I really liked Convince Me.*
- 8 It was a break from writing. I got to see some feedback.
- 9 It was neat to see a program that related all of my ideas.
- 10 Kinda fun. Helped define thinking. Learned about computer model, how "it" thinks people think. *Learned how to form arguments, convince the computer, helped make my thinking more precise and clear.*

Table 4.22. Written students' comments about the system, after reading about it.

Student Comment

- 1 Good for science classes esp. for elementary [to] high school since I've read "logical" step by step thinking isn't "natural" to our brains and takes training. I think the dotted/solid/converging lines diagram helps a lot in visually presenting logic, and should be emphasized in "*Convince Me*."...Teaches use of hypotheses and "scientific" thinking without human bias (esp. towards girls)...[but] a lot of redundancy—I think people can learn logical thinking faster with pen and paper.
- 2 It sounds like a great idea. Especially if one isn't too sure of one's hypothesis. It sounds like I could really use one of these programs, because I'm always wavering in my answers! It also helped to see one's options right there in front of you...I really need a program like this that systematically breaks down my thoughts...and gives me a clearer picture of where I stood.
- 3 It appears to be fairly objective. A good idea for papers. It's also a good start for AI (artificial intelligence). It's also unswayed by primacy bias—and can offer a non-emotional viewpoint. I'm convinced.
- 4 *Convince Me* seems like a worthy program that puts a lot of our thoughts together in a more cohesive way. A lot of times our ideas are jumbled, and we need some help to reach a central argument. To the extent that this is true would have been interesting to find out.
- 5 Seems just like any other word processing program, except some of the work is already done for you. But it's a good way to organize things, as they will appear easier to work with. I would probably figure out how to use it without Unit 3.
- 6 I think it would be useful in complex theories.
- 7 It seems to be a good way to get more organized about your ideas. It helps you to get a better overall picture about all the information given. But, I think it shouldn't be taken too seriously. The program seems to be very basic and inflexible. It is still kind of "primitive" for a computer program. It should only be used to give you a basic idea, an overview, of the information and your thoughts about the info.
- 8 The program seems very interesting. But the programming of the computer to decide whether the input of data by researchers is good or bad (by the ratings) seems so arbitrary. Of course, being a computer it can't reason on its own and decide if a piece of evidence is relevant or not. The program does sound very innovative and I would like to try it out. However, I would be very skeptical about the readings of the computer. The ratings that the computer has for each hypothesis reflects upon the beliefs of the designer and programmer of the software. I would be wary of the results.

- 9 The *Convince Me* program is a good way of testing one's own argument for believability. Though we may have many biases in favor of our argument the program bases its "belief" or "non-belief" on evidence and explanation given to it; therefore we can see the biases in our own reasoning and accordingly re-reason and explain our argument to try and make it more convincing. By changing the ratings on statements, or the parameter settings, we can see what is necessary to convince different people—for example: those with bias for evidence, those who are more skeptic and so on....[would] help to make one's own reasoning more clear by seeing one's own biases.
-

Discussion

The results reported above replicate the essential findings of Study 1 regarding hypotheses and evidence, and further show that statements can lie quite diversely about the two-dimensional hypothesis-/evidence-likeness space. But how critical are *Convince Me's* knowledge-eliciting interface and simulation-driven feedback for students' learning? The results of this study are very encouraging, in that they suggest that *Convince Me* makes its users better reasoners both while they employ it *and* when they are distal from it. The curriculum alone seems to help students (e.g., in the Written group) improve the coherence of their arguments, but this improvement is significantly lower than that evidenced by *Convince Me* users, and does not seem to (significantly) last as long (e.g., even through the post-test).

Convince Me seems to be a useful tool for structuring and revising arguments, in that when using the system (a) novices are more likely to reflect on and change the fundamental structure of their arguments when using the system (vs. Written students, who changed their ratings twice as often as their arguments), and (b) novices' beliefs were more in accord with the structures of their arguments, as evidenced by increased belief-activation correlations. Further, the software also yielded transfer in the latter case, in that belief-activation correlations for *Convince Me* users (a) did not significantly dip later when then did not have access to the

software (during the post-test), and (b) were higher than both their own pre-test and even the Written group's post-test. In contrast, the belief-activation correlation for the Written group rose less during the exercises, and was nonsignificantly higher than their pre-test performance. Since treatment time did not differ significantly between the Written and *Convince Me* groups (cf. Table 4.7), these differences seem attributable to the software (vs. time-on-task).

Convince Me does not seem to lead students to superficially revise their ratings to mimic ECHO; on the contrary, both the interface and feedback seem to help them articulate, connect, and revise arguments, and lead to a general improvement in the coherence between believability and argument structure. Thus, the *Convince Me* program appears to help students articulate and improve their arguments, even beyond the significant enhancement offered by the curriculum.

5. SUMMARY, DISCUSSION, AND CONCLUSIONS

Summary

In summary, Study 1 seriously (and empirically) questions the common implication that classifying evidence and hypotheses is reasonably straightforward once one formally studies science. It also further illustrates the generally positivistic stance held by most people (as was earlier observed by Schank & Ranney, 1991, etc.), and raises questions about how students might best be taught the "scientific method." This investigation is one of the few rigorous empirical pieces (and perhaps the only one) that shows that even experts in scientific reasoning—including those who have studied the distinction themselves—have difficulty discriminating data from theory. This seems true in their construct ratings, their inter-expert agreement, and their verbalizations. Further, although context facilitates the distinction, it by no means entirely obviates the difficulties; epistemological and semantic differences also cause disagreement about what constitutes "hypothesis" versus "evidence."

Convince Me and its associated paradigm improved novices' relatively deficient ability to discriminate between evidence and hypotheses, beyond the benefits of contextual embeddedness, even though the intervention employed here lasted only a few hours. For instance, novices demonstrated a much more negative post-test correlation between the constructs of hypothesis and evidence than was observed on pre-test measures; that is, they better differentiated the two constructs. This would appear to be an encouraging sign for developers of systems of this sort (cf. Cavalli-Sforza, Lesgold, & Weiner, 1992). Educators may not be able to successfully give students pat definitions of complex epistemic concepts like theory

and evidence, but they might aid their development through more sophisticated epistemological stages (e.g., Chandler, 1987). Although even experts disagree on the distinction between theory and data, *Convince Me* certainly makes novices respond more like experts during such epistemic categorizations.

Study 2 replicates the essential findings of Study 1 regarding hypotheses and evidence, and further empirically addresses the question of whether *Convince Me* is a tool and/or a training device to yield more coherent argumentation skills. Recall the question "Does the tool make its users (a) better reasoners while they employ it, (b) better reasoners even when they are distal from it, (c) both, or (d) neither?" Results indicate that the interface and feedback enhanced the students' learning, suggesting that the answer is "both." That is, while the curriculum and *Convince Me* manual represent materials that foster improved performance, when combined with the software and its associated feedback, students maintain more of their gains. *Convince Me* seems to help students structure and revise arguments, given that users' beliefs are more in accord with the structures of their arguments, and users are more likely to reflect on and change the fundamental structures of their arguments after getting feedback from the system. In other words, the full system may be viewed as both an effective "reasoner's workbench" tool *and* a learning environment that yields transfer to situations that are unsupported by the software and its attendant feedback.

Discussion

Can Reasoning be Modeled?

As discussed earlier (in the Introduction), several researchers have offered computational models of aspects of reasoning, many of which have met with limited success (e.g., DeJong & Mooney, 1986; Johnson, Krems, & Amra, 1994; Kintsch, 1988; Okada & Klahr, 1991; Pearl, 1988; Ram & Leake, 1991; Shultz & Lepper, 1992; Thagard & Millgram, in press). The investigations described here suggest that

explanatory reasoning can be modeled well (with ECHO) in a variety of contexts—from the more fictional to the more ecologically realistic, and using a range of post-hoc and predictive methods (e.g., involving both verbal protocols and text-based controversies). However, success was generally modulated by the amount of information available about an individual's knowledge base—and hence, indirectly, by context. Many simulations also rely on "primitive" categorizations of statements (e.g., "hypothesis," "evidence," "warrant," "backing," "premise," "conclusion," etc.), but the results described here show that people—even reasoning researchers themselves—may have difficulty making and agreeing on such epistemic categorizations (in concert with suggestions by others such as Hanson, 1958/1965 and Longino, 1990). Together, these findings suggest that to increase the chances of successfully modeling an individual's reasonings, researchers should not rely only on either representing the content presented *to* the individual or the experimenter's encodings of the individual's epistemic categorizations. Rather, researchers should explicitly elicit (and represent) background knowledge and epistemic views directly from the individual—as *Convince Me* does. To further improve the usefulness of such modeling, researchers should generally try to minimize the number of primitives and parameters (and hence the degrees of freedom) employed.

Can Reasoning be Improved?

As also discussed earlier, several researchers have illustrated difficulties that people have with formal and informal reasoning, and many others have identified useful teaching methods or tools to support reasoning (e.g., Brown & Campione, 1990; Eylon & Linn, in press; Hsi & Hoadley, 1994; Kuhn, 1993; Linn & Songer, 1993; Linn, Clement, & Pulos, 1983; Markman, 1979; Nickerson, Perkins, & Smith, 1985; Paolucci et al., 1995; Scardamalia & Bereiter, 1991; Schank & Ranney, 1992). These instructional methods include: teaching a few general evaluation principles,

reciprocal teaching, scaffolded integration, the use of computer learning environments to provide feedback or support argument development, and making deliberate use of context to leverage reasoning performance (e.g., via cognitive apprenticeship).

Results presented here illustrate how an "argument development environment" helps students reflect, revise, develop more sophisticated epistemic criteria, and build more coherent arguments. Such reflection and revision may also help students develop more integrated, dynamic views of science (cf. Linn & Songer, 1993; Eylon & Linn, in press), but this was not investigated here. The role of domain-independent feedback also calls for more investigation. The present studies suggest that it is effective, but further experiments are necessary to tease apart the precise utilities of *Convince Me's* argument interface from those of its feedback. While content and context-specific feedback are clearly important in reasoning instruction (e.g., Brown & Campione, 1990), figuring out what to believe in a wide variety of contexts is an important aspect of modern life. *Convince Me's* success in improving (and transferring) critical thinking should be viewed as complementary to (rather than a replacement for) more context-based instructional methods.

Finally, the present studies indicate that students' abilities to discriminate between the notions of hypotheses and evidence can indeed be improved, but even cognitive scientists cannot readily and reliably make such determinations. At the same time, many science textbooks (e.g., Giere, 1991; Starr & Taggart, 1984) start with descriptions of the distinction, or imply that the distinction is clear and apparent to experts. More accurate and honest portrayals of these constructs as fuzzy and dependent on context might even help students view science as a socially constructed, dynamic field—one that requires the continuous examination and revision of ideas rather than the memorization of disconnected "facts" (cf. Linn & Songer, 1993; Eylon & Linn, in press).

The Role of Context

Many researchers acknowledge that task-contexts influence reasoning, but some view these influences as unsystematic, or implicitly assume that knowledge and reasoning can be abstracted from the situation and applied (e.g., Piaget, 1970). If this were not possible, how would people transfer knowledge and reasoning to new situations? Do misconceptions reflect a depth of an individual's understanding of a situation's context, or do they reflect a lack of ability to use reasoning strategies effectively? Some argue that context effects are strong and that reasoning cannot be abstracted from the situation, and instruction should therefore make deliberate use of context (e.g., Brown, Collins, & Duguid, 1989; Brown & Campione, 1990; Lave & Wegner, 1991). In the extreme, the view that reasoning cannot be abstracted suggests that far-transfer (and perhaps even some near-transfer) is impossible.

Not surprisingly, the studies described here suggest that the answer lies somewhere between the two extremes—that is, both context and reasoning strategies are important. Subjects clearly brought extraneous information in their analyses of situations, even when given fictional texts that were generated to minimize the chance of this happening—and particularly when considering more realistic topics. Further, the effects of context on students' abilities to discriminate between the notions of evidence and hypothesis were generally (but not always) found to heighten the distinction. In concert with Johnson-Laird's (1983) theory of mental models, a context seems to reduce the space of scenarios in which a statement may serve. However, training with *Convince Me* made students' epistemic categorizations of both isolated *and* contextualized statements approach those of reasoning experts. Further, training made students behave more like experts in that they more strongly associated believability with evidence-likeness (versus hypothesis-likeness) both within and without a story context. Training also improved the degree to which students'

arguments reflected the strengths of their beliefs, in a variety of contexts. Together, these findings support the view that both general strategies and context are important to reasoning performance.

The Role of Technology

A number of computational models have usefully simulated aspects of reasoning (see the Introduction, and Can Reasoning Be Modeled? above). *Convince Me* has been very useful for the descriptive modeling of reasoning in that it automates the explication of both subjects' knowledge bases and their belief assessments. Pedagogically, technology has been useful in the general construction and management of arguments by others (e.g., Bereiter & Scardamalia, 1989; Hsi & Hoadley, 1994; Smolensky, Fox, King, & Lewis, 1988; VanLehn, 1985). It has also proved useful for scaffolding learning and reasoning in specific domains (e.g., Hartley, Byard, & Mallen, 1991; Eylon & Linn, in press; Ranney & Reiser, 1989; Reiser, Copen, Ranney, Hamid, & Kimberg, in press). Useful features of such environments include monitoring of student performance, prompts for students to reflect on their reasoning, feedback on predictions and performance, structured presentation of complex topic materials, and supportive management and editing tools. *Convince Me*, in association with the present curriculum, focuses on most of these features, while highlighting support for managing arguments and feedback to prompt reflection. Further, *Convince Me* is unique in that it uses a tested processing model to actually provide feedback on the plausibility of an argument's assertions—specifically for the benefit of students. The findings reported here support the view that carefully designed learning environments—for instance, ones that offers supportive management tools and encourage reflection via feedback—can significantly and quickly augment students' learning.

TEC-Based Modeling and Instruction

Convince Me was developed based on the instructional theory that students would benefit from a "reasoner's workbench" environment that both supports argument development and revision, and focuses reflection by providing simulation-based feedback on the consistency between one's conceptual environment and the strength of one's articulated beliefs. The primitives provided by the environment were also theory-driven (based on TEC and ECHO, which constrains the sorts of argument elements used to hypotheses, evidence, explanations, and contradictions) and are a distillation of primitives proposed by others (e.g., Toulmin, Rieke, & Janik, 1979; Paolucci et al., 1995). Simulation results and feedback were also theory-based (i.e., through the ECHO model). Our prior work supported the utility of TEC for modeling reasoning (e.g., Ranney et al., 1993; Schank & Ranney 1991, 1992), and the studies presented here supported the instructional utility of these theories in that *Convince Me* helped students revise and develop more coherent arguments.

What are the limitations of TEC-based modeling and instruction? We found ECHO largely successful in its modeling throughout a range of fictional-to-realistic contexts, but its success is modulated by (a) the amount of information it is provided about an individual's knowledge base (and hence, indirectly, by the context), as well as (b) the reliability with which hypotheses and evidence can be separated. Similarly, *Convince Me*'s domain-independence can be viewed as either a strength in that the system can be a general "reasoner's workbench," or a limitation in that it cannot give domain-specific feedback. The studies reported here show that even without domain-specific feedback, *Convince Me* can make its users better reasoners. However, other studies are needed to better understand the effectiveness of the argument interface versus ECHO's feedback. Further, the system could clearly be expanded to incorporate some content-knowledge and mechanisms for explaining its judgments to the students.

Future Directions

Adding More Representations (e.g., Diagrams) to *Convince Me*

Some have suggested that intensively training students (e.g., over several years) to methodically *analyze* and/or *diagram* "formal" arguments may be a way of getting them to reason more effectively. But formal arguments (depending on how one defines them) generally include much more in the way of logical implication and contradiction (as well as more direct sorts of classification; cf. Rosch, 1983). In contrast, "loose reasoning" (Ranney, in press) involves a complex of competing and supportive propositions that are generally neither absolutely implicative nor mutually exclusive in their relationships. Others have suggested that less formal philosophical arguments, but those still largely dealing with premises and conclusions, may provide a more compelling model for training students. But to suggest that even highly trained philosopher-logicians will be better at classifying premises and conclusions (in analogy to evidence and hypotheses), one must negate both some of the present (e.g., expert) findings, as well as the views of many philosophers. (E.g., many philosophical debates hinge on differing views of what are an argument's premises and conclusions—and what *kind* of premise or conclusion a particular proposition represents.)

Regarding diagrammatic representations, it may well be that some sort of graphical representation of one's argument will either improve one's ability to reason coherently or to classify propositions epistemically. Recent work by Ranney, Schank, and Diehl (1995) suggests that adding other linked representations to *Convince Me*—in particular, "global" representations such as argument diagrams and a listing of all of an argument's explanations and contradictions—will enhance its usefulness. Indeed, such a diagrammatic interface was recently added to the interface by the author (in Figure 5.1, see the top right, networked, "thermometer" icons), and

the ("Unit 3") *Convince Me* manual was extensively revised (with Christine Diehl) to reflect these new features (see Appendix H). However, individual aptitudes and predilections will likely determine whether such graphics will prove to be useful representations or interfering exercises. Both potential benefits and dangerous biases might arise from the use of either diagrammatic or textual/statement-based representations of arguments, and current studies focus on the utility of such representations (Diehl, Ranney, & Schank, 1995; cf. the usefulness various representations in the depiction of computer programs; e.g., Ranney & Reiser, 1989; Reiser, Ranney, Lovett, & Kimberg, 1989; Reiser et al., in press).

CM (big).v3.0 cubes

—Ratings— Add... Edit... Delete Rate... Rate All... Model's fit... Hide links

You		ECHO	Hypotheses:
5	5.4		H1. To make ice cubes freeze faster, use hot water, not cold water
6	5.3		H2. To make ice cubes freeze faster, use cold water, not hot water
7	5.6		H3. Water in the freezer should behave the same way as objects cooling to room temperature

You		ECHO	Evidence:
8	7		E1. The hotter something is, the longer it takes it to cool to room temperature
7	6.3		E2. Latisha's Mom found that hot water did freeze faster

Explanations: Explain... Delete Explanation

H3. Water in the freezer should behave the same way as objects cooling to room temperature "AND"
E1. The hotter something is, the longer it takes it to cool to room temperature

Explain(s) why: **"H2. To make ice cubes freeze faster, use cold water, not hot water"**

Contradictions: Conflict... Delete Conflict

H1. To make ice cubes freeze faster, use hot water, not cold water

Conflict(s) with: **"H2. To make ice cubes freeze faster, use cold water, not hot water"**

Help: E1. The hotter something is, the longer it takes it to cool to room temperature (click & drag node to rearrange graph) **File:**

Graph and simulation results:

All Explanations & Contradictions:

H1 explains E2
explains
H3 E1 jointly explain H2

H1 contradicts H2

Steps for using CONVINC ME:

1. Enter hypotheses and evidence
2. Enter explanations and contradictions
3. Rate the believability of your statements.
4. Run the ECHO simulation.
5. Compare your evaluations to ECHO's.
6. (optional) Make changes based on ECHO's feedback.

The correlation between your ratings and ECHO's evaluations is: 0.34 (mildly related).

The three most disparately rated statements are: H2, H1, H3, respectively (see boldened statements).

Your statement:

More of the hot water evaporates so there's less mass to freeze

Check all that apply:

Acknowledged fact or statistic

Observation or memory

One possible inference, opinion, or view

Some reasonable people might disagree

Select one:

Evidence E3 Reliability, if evidence? (from 1, poor, to 3, good)

Hypothesis H4

OK Cancel

Figure 5.1. Adding a belief to the ice cubes argument (cf. Figure 3.2) in response to *Convince Me's* feedback. This modified version of the software displays (a) an argument diagram (upper right) rather than merely displaying the "activational thermometer" icons in rows and columns, and (b) a listing of all explanations and contradictions (below the left side of the diagram).

Modeling Human Processing Limitations, and Other Model Modifications

Convince Me may also be improved by incorporating more "human" processing limitations into its modeling. Ranney (in press) notes that human reasoning is rarely as globally coherent as that of ECHO's memorially infallible connectionist algorithm, leading Hoadley et al. (1994) to (descriptively) model subjects' data with "WanderECHO"—i.e., ECHO with a limited attentional capacity (see Chapter 2). Explicitly contrasting ECHO's and WanderECHO's feedback may make students more aware of localities in their own reasoning (e.g., the momentary ignoring of discordant information; cf. Chinn & Brewer, 1993).

Others have argued for allowing students to specify the strength of their explanations and contradictions (which at present are largely set by TEC's principles and parameters). Similarly, at present ECHO allows only symmetric (versus directional) links between propositions, and some have argued that directional links might be more appropriate for certain kinds of arguments (e.g., deductive ones). Both of these suggested modifications would stretch the theory upon which ECHO is based (TEC) and would result in a model with more degrees of freedom (and perhaps too many to be useful), but they are certainly worthy of exploration.

Collaborative Work

An agreement among people as to what assumptions and hypotheses are embodied in a theory, what data are relevant, and even what the data are, is often difficult to obtain (e.g., Kuhn, 1993). *Convince Me* may prove useful in the future as a tool for collaboration, in ways similar to those of other environments such as CSILE (Scardamalia & Bereiter, 1991) and the *Convince Me*-inspired Interactive Multimedia Kiosk (Hsi & Hoadley, 1994), which support group dialectical processes. *Convince Me* helps students explicate their arguments, and thus could help students clarify and

share information with others, with the goal of understanding and incorporating others' theories and arguments in the service of building a group argument.

Social Versus "Scientific" Controversies, Typicality Studies

Future work might also focus more on "social" (vs. "scientific") controversies; for instance, how do knowledge structures differ between people who believe and disbelieve the Holocaust's genocide, or between those who differ on whether the moon landings were faked? Rosch-like typicality studies of epistemic categories could also be conducted to identify "prototypical" evidence and hypotheses, the graded structure of these "natural categories," and how they might vary across scientific, social, legal, and other contexts (e.g., Rosch, 1977, 1983). For instance, one group of subjects could generate exemplars of hypotheses or evidence, while others could rate the prototypicality of the generated hypotheses/evidence; reaction times for deciding whether an assertion is a hypothesis or a piece of evidence could also be collected, and assessed as to how they correlate with typicality ratings. Subjects could also list attributes of hypotheses and evidence, which could be used to identify a taxonomy for epistemic categories (e.g., with superordinate, basic, and subordinate levels, etc.)

Conclusions

The development of the *Convince Me* system partially grew from a desire to better descriptively model human belief assessment. In developing a more automated tool to elicit individuals' reasoning, however, I also created a pedagogical system that aids students in (a) articulating their theories and (b) revising such complexes of hypotheses and evidence in the face of feedback from the ECHO simulations. *Convince Me* seems to reduce the gap between the usual performance and potential competence of how one evaluates a complex, nondeductive, reasoning situation (e.g.,

via ratings of believability), partially by better articulating the elements and relationships that spawn such evaluations (Ranney et al., 1995). Results further indicate that the system may be viewed as both a tool to articulate and refine one's thinking *and* a training/learning system that fosters the transfer of such articulation to situations in which the software is not available.

REFERENCES

- Andersen, S. K., Jensen, F. V., Olesen, K. G., & Jensen, F. (1989). *HUGIN: A shell for building Bayesian belief universes for expert systems* [computer program]. Aalborg, Denmark: HUGIN Expert Ltd.
- Anderson, J.R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.
- Anderson, J.R. (1985). *Cognitive psychology and its implications*. New York, NY: W.H. Freeman and Company.
- Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological review*, *94*, 192-210.
- Austin, L.B., & Shore, B.M. (1993). Concept mapping of high and average achieving students and experts. *European Journal For High Ability*, *4*, 180-195.
- Bar-On, E. (1991). *Mental capacity and locally-coherent views: Towards a unifying theory*. Draft. Department of Science Education, Technion, Israel.
- Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, England: University Press.
- Bereiter, C., & Scardamalia. (1989). When weak explanations prevail. (Commentary on P. Thagard's "Explanatory Coherence," same issue.) *Behavioral and Brain Science*, *12*, 468-469.
- Boole, G. (1854) *An investigation of the laws of thought*. London: Walton and Maberly.

- Bradshaw, G. L., & Anderson, J.R. (1982). Elaborative encoding as an explanation of levels of processing. *Journal of Verbal Learning and Verbal Behavior*, *21*, 165-174.
- Bransford, J.D., & Johnson, M.K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, *11*, 717-726.
- Brewer, W., & Chinn, C. (1991). Entrenched beliefs, inconsistent information, and knowledge change. In L. Birnbaum (Ed.), *Proceedings of the International Conference on the Learning Sciences*, 67-73. Charlottesville, VA: Association for the Advancement of Computing in Education (AACE).
- Brewer, W.F., & Treyens, J. C. (1981). The role of schemata in memory for places. *Cognitive Psychology*, *13*, 207-230.
- Brown, A. L., & Campione, J. C. (1990). Communities of learning and thinking: A context by any other name. In D. Kuhn (Ed.), *Developmental Perspectives on Teaching and Learning Thinking Skills* [Special Issue]. *Contributions to Human Development*, *21*, 108-126.
- Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *33*, 32-42.
- Burbules, N. (1992). The virtues of reasonableness. In M. Buchmann & R. Floden (Eds.), *Philosophy of Education 1991*, 215-224. Normal, IL: Philosophy of Education Society.

- Burbules, N. (1995). Reasonable doubt: Toward a postmodern defense of reason as an educational aim. In W. Kohli (Ed.), *Critical Conversations in Philosophy of Education*, 82-102. New York, NY: Routledge.
- Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.
- Carlock, M. (1990). *Learning coherence through MacECHO IFE: An analysis of coherence*. Second Year Project Report, Univ. of California, Berkeley, Graduate School of Education.
- Case, R. (1974). Structures and strictures: Some functional limitations on the course of cognitive growth. *Cognitive Psychology*, 6, 544-573.
- Cavalli-Sforza, V., Lesgold, A.M., & Weiner, A.W. (1992). Strategies for contributing to collaborative arguments. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, 755-760. Hillsdale, NJ: Erlbaum.
- Cavalli-Sforza, V., Moore, J., & Suthers, D. (1993). Helping students articulate, support, and criticize scientific explanations. In *Proceedings of the World Conference on Artificial Intelligence in Education*, 113-120. Charlottesville, NC: Association for the Advancement of Computing in Education.
- Chandler, M. (1987) The Othello Effect: Essay on the emergence and eclipse of skeptical doubt. *Human Development*, 30, 137-159.
- Cheng, P., & Holyoak, K. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, 17, 391-416.
- Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152.

- Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63*, 1-49.
- Cohen, L. J. (1989). Two problems for the explanatory coherence theory of acceptability. (Commentary on P. Thagard's "Explanatory Coherence," same issue.) *Behavioral and Brain Science, 12*, 471.
- DeJong, G. & Mooney, R. (1986). Explanation-based learning: An alternative view. *Machine Learning, 1*, 145-176.
- Diehl, C., Ranney, M., & Schank, P. (1995, April). *Multiple representations for improving scientific thinking*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- diSessa, A. (1983). Phenomenology and the evolution of intuition. In D. Gentner & A. Stevens (Eds.), *Mental models*, 15-33. Hillsdale, NJ: Lawrence Erlbaum.
- diSessa, A.A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10*, 105-225.
- Dreyfus, H. (1992). *What computers still can't do*. Cambridge, MA: MIT Press.
- Ericsson. K.A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
- Eylon, B., & Linn, M. (in press). Models and integration activities in science education. In E. Bar-On, B. Eylon, & Z. Schertz (Eds.), *Designing intelligent learning environments: From cognitive analysis to computer implementation*. Norwood, NJ: Ablex.

- Feldman, J. A. (1989). What's in a link? (Commentary on P. Thagard's "Explanatory Coherence," same issue.) *Behavioral and Brain Science*, 12, 474-475.
- Feyerabend, P. (1978). *Against method: Outline of an anarchist theory of knowledge*. London: Verso.
- Fiske, S., & Taylor, S. (1984). *Social cognition*. New York: Random House.
- Freedman, E.G. (1992). Understanding scientific controversies from a computational perspective: The case of latent learning. In R.N. Giere (Ed.), *Cognitive models of science*, 310-335. Minneapolis, MN: University of Minnesota Press (In the Minnesota Studies of Philosophy of Science series; Vol. 15.)
- Friedler, Y., Nachmias, R., & Songer, N. (1989). Teaching scientific reasoning skills: A case study of a microcomputer-based curriculum. *School Science and Mathematics*, 89, 1, 59-67.
- Gabrys, G. (1989). HEIDER: A simulation of attitude consistency and attitude change. In S. Ohlsson (Ed.), *Aspects of cognitive conflict and cognitive change*. Univ. of Pittsburgh, Technical Report, KUL-89-04.
- Garnham, A. & Oakhill, J. (1944). *Thinking and reasoning*. Cambridge, MA: Blackwell.
- Gentner, D. & Grudin, J. (1985). The evolution of mental metaphors in psychology: A ninety-year retrospective. *American Psychologist*, 40, 181-192.
- Gentner, D., & Stevens, A. L. (Eds.) (1983). *Mental models*. Hillsdale, NJ: Lawrence Erlbaum.

- Giere, R.N. (1991). *Understanding scientific reasoning*. New York: Holt, Rinehart, and Winston.
- Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, 98, 2, 254-267.
- Givon, T. (1991, August). *Coherence: Toward a cognitive model*. Paper presented at the First Annual Meeting of the Society for Text and Discourse, Chicago, IL.
- Goodman, N. (1954). *Fact, Fiction, and Forecast* (1st edition). Cambridge, Massachusetts: Harvard University Press (4th edition, 1983).
- Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.). *Syntax and Semantics*, vol 3: Speech Acts. New York, NY: Seminar Press.
- Hammer, D. (1994). Epistemological beliefs in introductory physics. *Cognition and Instruction*, 12, 151-183.
- Hanson, N.R. (1965). *Patterns of discovery: An inquiry into the conceptual foundations of science*. London: Cambridge University Press. (Original work published in 1958.)
- Harman, G. (1989). Competition for evidential support. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, 220-226. Hillsdale, NJ: Erlbaum.
- Hartley, J. R., Byard, M. J., & Mallen, C. L. (1991). Qualitative modeling and conceptual change in science students. In L. Birnbaum (Ed.), *Proceedings of the International Conference on the Learning Sciences*, 222-230. Charlottesville, VA: AACE.

- Hastie, R. (1980). Memory for behavioral information that confirms or contradicts a personality impression. In R. Hastie, T.M. Ostrom, E.B. Ebbesen, R.S. Wyer, Jr., D.L. Hamilton, and D.E. Carlston (Eds.) *Person memory: The cognitive basis of social perception*, 1-53. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hastie, R., & Carlston, D. (1980). Theoretical issues in person memory. In R. Hastie, T.M. Ostrom, E.B. Ebbesen, R.S. Wyer, Jr., D.L. Hamilton, and D.E. Carlston (Eds.) *Person memory: The cognitive basis of social perception*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hoadley, C., Ranney, M., & Schank, P. (1994). WanderECHO: A connectionist simulation of limited coherence. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, 421-426, Hillsdale, NJ: Erlbaum.
- Holland, J. (1992). *Adaption in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Cambridge, MA: MIT Press.
- Holland, J., Holyoak, K., Nisbett, R., & Thagard, P. (1989). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.
- Holyoak, K., & Spellman, B. (1993). Thinking. *Annual Review of Psychology*, 44, 265-315.
- Hsi, S. & Hoadley, T. (1994, April). *An interactive multimedia kiosk as a tool for collaborative discourse, reflection, and assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

- Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.
- Johnson, T., Krems, J., & Amra, N. (1994). A computational model of human abductive skill and its acquisition. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, 463-468. Hillsdale, NJ: Erlbaum.
- Johnson, T. & Smith, J. (1991). A framework for opportunistic abductive strategies. *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, 760-764. Hillsdale, NJ: Erlbaum.
- Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.
- Johnson-Laird, P. (1988). *The computer and the mind: An introduction to cognitive science*. Cambridge, MA: Harvard University Press.
- Keller, E.F. (1985). *Reflections on gender and science*. New Haven, CT: Yale University Press.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163-182.
- Kintsch, W. (1992). How readers construct situation models for stories: The role of syntactic cues and causal inferences. In A. F. Healy, S. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes*, vol. 2. Hillsdale, NJ: Erlbaum.
- Klahr, D. & Dunbar, K. (1988). Dual search space during scientific reasoning. *Cognitive Science*, 12, (1), 1-48.

- Kolers, P. & Smythe, W. (1984). Symbol manipulation: Alternatives to the computational view of mind. *Journal of Verbal Learning and Verbal Behavior*, 23, 289-314.
- Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, 96, 674-689.
- Kuhn, D. (1993). Connecting scientific and informal reasoning. *Merrill-Palmer Quarterly-Journal of Developmental Psychology*, 39, 1, 74-103.
- Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific thinking skills*. Orlando, FL: Academic Press.
- Laird, J., Newell, A., & Rosenbloom, P. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33, 1-64.
- Laird, J., Rosenbloom, P., & Newell, A. (1986). Chunking in SOAR: The anatomy of a general learning mechanism. *Machine Learning*, 1, 11-46.
- Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: The University of Chicago Press.
- Lave, J. & Wegner, E. (1991). *Situated learning: Legitimate peripheral participation*. New York: Cambridge University Press.
- Linn, M. C. (1990). What constitutes scientific thinking? [Book review, Kuhn, et al., *The development of scientific thinking*, San Diego, CA: Academic Press, 1988]. *Contemporary Psychology*, 35(1), 16-17.
- Linn, M. C., Clement, C., & Pulos, S. (1983). Is it formal if it's not physics? *Journal of Research in Science Teaching*, 20(8), 755-770.

- Linn, M., & Songer, N. (1993). How do students make sense of science? *Merrill-Palmer Quarterly-Journal of Developmental Psychology*, 39, 47-73.
- Lipman, M. (1985). Thinking skills fostered by philosophy for children. In J. W. Segal, S. F. Chipman, & R. Glaser (Eds.), *Thinking and Learning Skills, Volume 1: Relating Instruction to Research*, pp. 83-108. Hillsdale, NJ: Erlbaum.
- Lipman, M. (1991). *Thinking in education*. Cambridge, UK: Cambridge University Press.
- Longino, H. E (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton, NJ: Princeton University Press.
- Mangan, B., & Palmer, S. (1989). New science for old. (Commentary on P. Thagard's "Explanatory Coherence," same issue.) *Behavioral and Brain Sciences*, 12, 480-482.
- Markman, E.M. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. *Child Development*, 59, 643-655.
- McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, 2, 6, 387-395.
- McKoon, G., & Ratcliff, R. (1980). Priming in item recognition: The organization of propositions in memory for text. *Journal of Verbal Learning and Verbal Behavior*, 19, 369-386.
- Merrill, D.C., & Reiser, B.J. (1994). Scaffolding effective problem solving strategies in interactive learning environments. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, 629-634. Hillsdale, NJ: Erlbaum.

- Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227-234.
- Miller, L. C., & Read, J. S. (1991). On the coherence of mental models of persons and relationships: a knowledge structure approach. In F. Fincham and G.J.O. Fletcher (Eds.) *Cognition in close relationships*, 69-99. Hillsdale, NJ: Lawrence Erlbaum.
- Nersessian, N. J. (1989). Conceptual change in science and in science education. *Synthese*, *80*, 163-183.
- Nickerson, R., Perkins, D., & Smith, E. (1985). *The teaching of thinking*. Hillsdale, NJ: Erlbaum.
- Okada, T., & Klahr, D. (1991). Searching an hypothesis space when reasoning about buoyant forces: The effect of feedback. *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, 842-846. Hillsdale, NJ: Erlbaum.
- Osherson, D. (1990). Judgment. In D. Osherson & E. Smith (Eds.), *An invitation to cognitive science: Thinking* (vol. 3), 55-87. Cambridge, MA: MIT Press.
- Paolucci, M., Suthers, D., & Weiner, A. (1995). Belvedere: Stimulating student's critical discussion. *Human Factors in Computing Systems CHI '95 Conference Companion*, 123-124. New York, NY: Association for Computing Machinery.
- Papineau, D. (1989). Probability and normality. (Commentary on P. Thagard's "Explanatory Coherence," same issue.) *Behavioral and Brain Science*, *12*, 484-485.

- Pearl, J. (1986). A constraint-propagation approach to probabilistic reasoning. In L.N. Kanal & J.F. Lemmer (Eds.), *Uncertainty in artificial intelligence*, 357-369. Amsterdam: North-Holland.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo: Morgan Kaufmann.
- Pennington, N. & Hastie, R. (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14, 521-533.
- Perkins, D. N., Allen, R., & Hafner, J. (1983). Difficulties in everyday reasoning. In W. Maxwell (Ed.), *Thinking: The Expanding Frontier*, Philadelphia, PA: The Franklin Institute Press.
- Perkins, D. N. (1985). Postprimary education has little impact on informal reasoning. *Journal of Educational Psychology*, 77, 562-571.
- Piaget, J. (1970). Piaget's theory. In P.R. Mussen (Ed.), *Carmichael's handbook of child psychology* (3rd ed.), 703-732. New York: Wiley.
- Popper, K.R. (1978). *Conjectures and refutations* (rev. ed.). London: Routledge & K. Paul.
- Quine, W. & Ullian, J. (1970). *The web of belief*. New York, NY: Random House.
- Ram, A. & Leake, D. (1991). Evaluation of explanatory hypotheses. *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, 867-871. Hillsdale, NJ: Erlbaum.

- Ranney, M. (1988). Changing naive conceptions of motion (Doctoral dissertation, University of Pittsburgh, Learning Research and Development Center, 1987). *Dissertation Abstracts International*, 49, 1975B.
- Ranney, M. (1994a). Assessing and contrasting formal and informal/experiential understandings of trajectories. In the G. Marks (Ed.) *Proceedings of the International Symposium on Mathematics/Science Education and Technology*, 142-146, Charlottesville, VA: AACE.
- Ranney, M. (1994b). Relative consistency and subjects' "theories" in domains such as naive physics: Common research difficulties illustrated by Cooke & Breedin. *Memory & Cognition*, 4, 494-502.
- Ranney, M. (in press). Explorations in explanatory coherence. In E. Bar-On, B. Eylon, & Z. Schertz (Eds.), *Designing intelligent learning environments: From cognitive analysis to computer implementation*. Norwood, NJ: Ablex.
- Ranney, M., & Reiser, B.J. (1989). Reasoning and explanation in an intelligent tutor for programming. In G. Salvendy & M.J. Smith (Eds.), *Designing and using human-computer interfaces and knowledge based systems*, 88-95. New York: Elsevier Science Publishers.
- 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., & Diehl, C. (1995). Competence versus performance in critical reasoning: reducing the gap by using *Convince Me*. *Psychology Teaching Review*, 4, 2, 151-164.

- 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. [An updated version will appear in B.H. Kwasnik (Ed.) (in press), *Advances in classification research* (ASIS Monograph Series), Medford, NJ: Learned Information]
- Ranney, M., Schank, P., Mosmann, A., & Montoya, G. (1993). Dynamic explanatory coherence with competing beliefs: Locally coherent reasoning and a proposed treatment. In T.-W. Chan (Ed.), *Proceedings of the International Conference on Computers in Education: Applications of Intelligent Computer Technologies*, 101-106.
- Ranney, M., Schank, P., & Ritter, C. (1992, January). *Studies of explanatory coherence using text, discourse, and verbal protocols*. Paper presented at the Third Annual Winter Text Conference, Jackson, Wyoming.
- Ranney, M., Schank, P., Ritter, C., & Carlock, M. (1991, March). *Descriptive and prescriptive studies of explanatory coherence*. Paper presented at the Third Biennial Cognition and Instruction Workshop, Pittsburgh, PA.
- Ranney, M., & Thagard, P. (1988). Explanatory coherence and belief revision in naive physics. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, 426-432. Hillsdale, NJ: Erlbaum. (Also appears as Report No. SE-050-095, ERIC Document Reproduction Service No. ED 301 407; pp. 1-15)
- Read, S.J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65, 429-447.

- Reiser, B.J., Copen, W.A., Ranney, M., Hamid, A., & Kimberg, D.Y. (in press). Cognitive and motivational consequences of tutoring and discovery learning. *Cognition and Instruction*.
- Reiser, B.J., Kimberg, D.Y., Lovett, M.C., & Ranney, M. (1992). Knowledge representation and explanation in GIL, an intelligent tutor for programming. In J.H. Larkin & R.W. Chabay (Eds.), *Computer assisted instruction and intelligent tutoring systems: Shared goals and complementary approaches*, 111-149. Hillsdale, NJ: Erlbaum.
- Reiser, B.J., Ranney, M., Lovett, M.C., & Kimberg, D.Y. (1989). Facilitating students' reasoning with causal explanations and visual representations. In D. Bierman, J. Breuker, & J. Sandberg (Eds.), *Proceedings of the Fourth International Conference on Artificial Intelligence and Education*, 228-235. Springfield, VA: IOS.
- Rips, L. (1983). Cognitive processes in propositional reasoning. *Psychological Review*, 90, 38-71.
- Ritter, C. (1991). *Thinking about ECHO*. Unpublished master's project, University of California, Berkeley.
- Rosch, E. (1977). Human categorization. In N. Warren (Ed.), *Studies in cross-cultural psychology*, 3-49. New York: Academic Press.
- Rosch, E. (1983). Prototype classification and logical classification: The two systems. In E.K. Scholnick (Ed.), *New trends in conceptual representation: Challenges to Piaget's theory?*, 73-86. Hillsdale, NJ: Erlbaum.

- Rumelhart, D. E., McClelland, J. R. & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (vols. 1 & 2). Cambridge, MA: MIT Press.
- Russell, S., & Wefald, E. (1991). *Do the right thing*. Cambridge, MA: MIT Press.
- Saxe, G. B. (1988). Candy selling and math learning. *Educational Researcher*, 17 (6), 14-21.
- 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., Hoadley, C., Dougery, K., Neff, J., & Ranney, M. (1993). *The ECHO Educational Program (EEP) Coherent Reasoning Curriculum for Convince Me* [Computer Program Manual and Curriculum]. University of California, Berkeley, Graduate School of Education.
- Schank, P., Linn, M., & Clancy, M. (1993). Supporting Pascal programming with an on-line template library and case studies. *International Journal of Man-Machine Studies*, 38, 1031-1048.
- 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*, 892-897. Hillsdale, NJ: Erlbaum.
- Schank, P., & Ranney, M. (1992). Assessing explanatory coherence: A new method for integrating verbal data with models of on-line belief revision. *Proceedings of*

the Fourteenth Annual Conference of the Cognitive Science Society, 599-604.
Hillsdale, NJ: Erlbaum.

Schank, P., & Ranney, M. (1993). Can reasoning be taught? *Educator*, 7 (1), 16-21.
[Special issue on cognitive science and education].

Schank, P., & Ranney, M. (1995). Improved reasoning with *Convince Me*. *Human Factors in Computing Systems CHI '95 Conference Companion*, 276-277. New York, NY: Association for Computing Machinery.

Schank, P., Ranney, M., & Hoadley, C. (1995). *Convince Me* [Computer program and manual]. In J.R. Jungck, N. Peterson, & J.N. Calley (Eds.), *The BioQUEST Library*. College Park, MD: Academic Software Development Group, University of Maryland.

Schank, P., Ranney, M., Hoadley, C., Diehl, C., & Neff, J. (1994). A reasoner's workbench for improving scientific thinking: Assessing *Convince Me*. In the G. Marks (Ed.) *Proceedings of the International Symposium on Mathematics/Science Education and Technology*, 237. Charlottesville, VA: AACE.

Schank, P., and Rowe, L. (1993). The design and assessment of a hypermedia course on semiconductor manufacturing. *Journal of Educational Multimedia and Hypermedia*, 2 (3), 299-320.

Searle, J. (1990). Is the brain's mind a computer program? *Scientific American*, 262 (1), 25-31.

- Shultz, T., & Lepper, M. (1992). A constraint satisfaction model of cognitive dissonance phenomena. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, 462-467. Hillsdale, NJ: Erlbaum.
- Sime, J. (1993). Modeling a learner's multiple models with basin belief networks. In P. Brna, S. Ohlsson, & H. Pain (Eds.), *Proceedings of AI-Ed '93: World Conference on Artificial Intelligence in Education*, 426-432. Charlottesville, VA: AACE.
- Simon, H. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, 69, 99-118.
- Slovic, P. (1990). Choice. In D. Osherson & E. Smith (Eds.), *An invitation to cognitive science: Thinking*, 89-115. Cambridge, MA: MIT Press.
- Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11, 1-74.
- Smolensky, P., Fox, B., King, R., & Lewis, C. (1988). Computer-aided reasoned discourse or, how to argue with a computer. In R. Guindon (Ed.), *Cognitive Science and its Applications for Human-Computer Interaction*, 109-162. Hillsdale, NJ: Erlbaum.
- Starr, C. & Taggart, R. (1984). *Biology: The unity and diversity of life*. Belmont, CA: Wadsworth Publishing Company.
- Stich, S. (1990). Rationality. In D. Osherson & E. Smith (Eds.), *An invitation to cognitive science: Thinking*, 173-196. Cambridge, MA: MIT Press.
- Suthers, D., Weiner, A., Connelly, J., & Paolucci, M. (1995). Belvedere: Engaging students in critical discussion of science and public policy issues. *Proceedings of*

- AI-Ed '95: Seventh World Conference on Artificial Intelligence in Education*, 266-273. Charlottesville, NC: Association for the Advancement of Computing in Education.
- Tash, J. (1994). Formal rationality and limited agents. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, 854-858. Hillsdale, NJ: Erlbaum.
- Tash, J., & Russell, S. (1994). Control strategies for a stochastic planner. *Proceedings of the Twelfth National Conference on Artificial Intelligence*. Cambridge, 1079-1085. Cambridge, MA: AAAI Press/MIT Press.
- Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12, 435-502.
- Thagard, P. (1991a). The dinosaur debate: Explanatory coherence and the problem of competing hypotheses. In J. Pollock and R. Cummins (Eds.), *Philosophy and AI: Essays at the interface*, 279-300. Cambridge, MA: MIT Press/Bradford Books.
- Thagard, P. (1991b). Philosophical and computational models of explanation. *Philosophical Studies*, 64, 87-104.
- Thagard, P.R. (1992). *Conceptual revolutions*. Princeton, NJ: Princeton University Press.
- Thagard (forthcoming). *Coherence*. Unpublished manuscript. Philosophy Department, University of Waterloo, Waterloo, Ontario.

- Thagard, P. (in press). Probabilistic networks and explanatory coherence. In P. O'Rorke & J. Josephson (Eds.) *Automated abduction: Inference to the best explanation*. Menlo Park, CA: AAAI Press.
- Thagard, P., & Millgram, E. (in press). Inference to the best plan: A coherence theory of decision. In D. Leake & A. Ram (Eds.), *Goal-directed learning*. Cambridge, MA: MIT Press.
- Thagard, P. and Nelson, G. (1988). *ECHO* [computer program], Common Lisp version. Princeton University.
- Thorndike, E.L. (1922). *The psychology of arithmetic*. New York: Macmillan.
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge, England: Cambridge University Press.
- Toulmin, S. E., Rieke, R., & Janik, A. (1979). *An introduction to reasoning*. New York: Macmillan.
- Trabasso, T., van den Broek, P., & Suh, S. (1989). Logical necessity and transitivity of causal relations in stories. *Discourse Processes*, 12, 1-25.
- Tversky, A., & Kahneman, D. (1976). Judgments under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Tweney, R.D., Doherty, M.E., & Mynatt, C.R. (Eds.). (1981). *On scientific thinking*. New York: Columbia University Press.
- VanLehn, K. (1985). *Theory reform caused by an argumentation tool*. Xerox Palo Alto Research Center Report P85-00102.

- Verbeurgt, K. & Thagard, P. (forthcoming). *The computational complexity of coherence and partial constraint satisfaction*. Unpublished manuscript. Philosophy Department, University of Waterloo, Waterloo, Ontario.
- Villano, M. (1992). Probabilistic student models: Bayesian belief networks and knowledge space theory. *Proceedings of the 2nd International Conference on Intelligent Tutoring Systems, Lecture Notes in Computer Science, 608*. Berlin: Springer-Verlag.
- Wason, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273-281.
- Wason, P. C. (1977). 'On the failure to eliminate hypotheses...'—a second look. In P. N. Johnson-Laird, & P. C. Wason (Eds.), *Readings in Cognitive Science*. NY: Cambridge University.
- Wason, P.C., & Johnson-Laird, P.N. (1972). *Psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University Press.
- Winograd, T., & Flores, F. (1987). *Understanding computers and cognition*. Reading, MA: Addison Wesley.
- Woodsworth, F., & Sells, S. (1935). An atmospheric effect in formal syllogistic reasoning. *Journal of Experimental Psychology, 18*, 451-460.
- Wyer, R.S. Jr., & Srull, T. K. (1989). *Memory and cognition in its social context*. Hillsdale, NJ: Lawrence Erlbaum Associates.

APPENDIX A: Pre-Test

Name _____

Date _____

Problem 1

Give definitions for the following:

- a) hypothesis
- b) evidence
- c) fact
- d) explanation
- e) contradiction
- f) theory
- g) argument
- h) confirmation bias
- i) disconfirmation
- j) recency bias
- k) primacy bias

Problem 2

Based **your** view and knowledge of the world, for each of the following statements please:

1. Rate (circle) how good an example of a *hypothesis* you think the statement is,
2. Rate (circle) how good an example of a piece of *evidence* you think the statement is,
3. Explain (briefly, in writing) why you gave the hypothesis and evidence ratings you did, and
4. Rate (circle) how strongly you *believe* the statement.

[Note: statements b,d,f,h j l n, & p were not included in Study 1]

a) All wine is made from grapes.

definitely <u>not</u> hypothesis	1	2	3	4	5	6	7	8	9	definitely hypothesis
-------------------------------------	---	---	---	---	---	---	---	---	---	--------------------------

definitely <u>not</u> evidence	1	2	3	4	5	6	7	8	9	definitely evidence
-----------------------------------	---	---	---	---	---	---	---	---	---	------------------------

why (explain): _____

completely disbelieve/reject	1	2	3	4	5	6	7	8	9	completely believe/accept
---------------------------------	---	---	---	---	---	---	---	---	---	------------------------------

b) Some dogs have an aggressive disorder in which they bark more, growl more, bite more, and have higher blood pressure and heart rate than other dogs do.

definitely <u>not</u> hypothesis	1	2	3	4	5	6	7	8	9	definitely hypothesis
-------------------------------------	---	---	---	---	---	---	---	---	---	--------------------------

definitely <u>not</u> evidence	1	2	3	4	5	6	7	8	9	definitely evidence
-----------------------------------	---	---	---	---	---	---	---	---	---	------------------------

why (explain): _____

completely disbelieve/reject	1	2	3	4	5	6	7	8	9	completely believe/accept
---------------------------------	---	---	---	---	---	---	---	---	---	------------------------------

1 2 3 4 5 6 7 8 9

c) Gravity exists in other galaxies.

definitely	<u>not</u>								definitely
hypothesis			neutral					hypothesis	
1	2	3	4	5	6	7	8	9	

definitely	<u>not</u>								definitely
evidence			neutral					evidence	
1	2	3	4	5	6	7	8	9	

why (explain): _____

completely									completely
disbelieve/reject			neutral					believe/accept	
1	2	3	4	5	6	7	8	9	

d) Lack of a chemical causes an aggressive disorder in dogs.

definitely	<u>not</u>								definitely
hypothesis			neutral					hypothesis	
1	2	3	4	5	6	7	8	9	

definitely	<u>not</u>								definitely
evidence			neutral					evidence	
1	2	3	4	5	6	7	8	9	

why (explain): _____

completely									completely
disbelieve/reject			neutral					believe/accept	
1	2	3	4	5	6	7	8	9	

e) Gravity exists on Earth.

definitely	<u>not</u>								definitely
hypothesis			neutral					hypothesis	
1	2	3	4	5	6	7	8	9	

definitely	<u>not</u>								definitely
evidence			neutral					evidence	
1	2	3	4	5	6	7	8	9	

why (explain): _____

completely									completely
disbelieve/reject			neutral					believe/accept	
1	2	3	4	5	6	7	8	9	

f) Some researchers trained one group of aggressive-disorder dog owners to treat their dogs firmly yet lovingly.

definitely <u>not</u> hypothesis	1	2	3	neutral	4	5	6	7	definitely hypothesis	8	9
definitely <u>not</u> evidence	1	2	3	neutral	4	5	6	7	definitely evidence	8	9
why (explain): _____											
completely disbelieve/reject	1	2	3	neutral	4	5	6	7	completely believe/accept	8	9

g) President John F. Kennedy was assassinated.

definitely <u>not</u> hypothesis	1	2	3	neutral	4	5	6	7	definitely hypothesis	8	9
definitely <u>not</u> evidence	1	2	3	neutral	4	5	6	7	definitely evidence	8	9
why (explain): _____											
completely disbelieve/reject	1	2	3	neutral	4	5	6	7	completely believe/accept	8	9

h) Some researchers found that training dog owners to treat their dogs firmly yet lovingly relieved symptoms of aggressive disorder in their dogs.

definitely <u>not</u> hypothesis	1	2	3	neutral	4	5	6	7	definitely hypothesis	8	9
definitely <u>not</u> evidence	1	2	3	neutral	4	5	6	7	definitely evidence	8	9
why (explain): _____											
completely disbelieve/reject	1	2	3	neutral	4	5	6	7	completely believe/accept	8	9

i) Abraham Lincoln said that Ross Perot would lose in 1992.

definitely <u>not</u> hypothesis	1	2	3	4	5	6	7	8	9	definitely hypothesis
-------------------------------------	---	---	---	---	---	---	---	---	---	--------------------------

definitely <u>not</u> evidence	1	2	3	4	5	6	7	8	9	definitely evidence
-----------------------------------	---	---	---	---	---	---	---	---	---	------------------------

why (explain): _____

completely disbelieve/reject	1	2	3	4	5	6	7	8	9	completely believe/accept
---------------------------------	---	---	---	---	---	---	---	---	---	------------------------------

j) Some researchers think dogs get an aggressive disorder when their owners treat them poorly.

definitely <u>not</u> hypothesis	1	2	3	4	5	6	7	8	9	definitely hypothesis
-------------------------------------	---	---	---	---	---	---	---	---	---	--------------------------

definitely <u>not</u> evidence	1	2	3	4	5	6	7	8	9	definitely evidence
-----------------------------------	---	---	---	---	---	---	---	---	---	------------------------

why (explain): _____

completely disbelieve/reject	1	2	3	4	5	6	7	8	9	completely believe/accept
---------------------------------	---	---	---	---	---	---	---	---	---	------------------------------

k) Birds evolved from animals that lived in trees.

definitely <u>not</u> hypothesis	1	2	3	4	5	6	7	8	9	definitely hypothesis
-------------------------------------	---	---	---	---	---	---	---	---	---	--------------------------

definitely <u>not</u> evidence	1	2	3	4	5	6	7	8	9	definitely evidence
-----------------------------------	---	---	---	---	---	---	---	---	---	------------------------

why (explain): _____

completely disbelieve/reject	1	2	3	4	5	6	7	8	9	completely believe/accept
---------------------------------	---	---	---	---	---	---	---	---	---	------------------------------

l) Other researchers found that a chemical relieved symptoms of aggressive disorder in dogs.

definitely <u>not</u> hypothesis	1	2	3	neutral	4	5	6	7	definitely hypothesis	8	9
definitely <u>not</u> evidence	1	2	3	neutral	4	5	6	7	definitely evidence	8	9
why (explain): _____											
completely disbelieve/reject	1	2	3	neutral	4	5	6	7	completely believe/accept	8	9

m) Approximately three-quarters of the surface of the Earth is covered by water.

definitely <u>not</u> hypothesis	1	2	3	neutral	4	5	6	7	definitely hypothesis	8	9
definitely <u>not</u> evidence	1	2	3	neutral	4	5	6	7	definitely evidence	8	9
why (explain): _____											
completely disbelieve/reject	1	2	3	neutral	4	5	6	7	completely believe/accept	8	9

n) Abuse causes an aggressive disorder in dogs.

definitely <u>not</u> hypothesis	1	2	3	neutral	4	5	6	7	definitely hypothesis	8	9
definitely <u>not</u> evidence	1	2	3	neutral	4	5	6	7	definitely evidence	8	9
why (explain): _____											
completely disbelieve/reject	1	2	3	neutral	4	5	6	7	completely believe/accept	8	9

o) All humans on Earth are dead at this moment.

definitely <u>not</u> hypothesis	1	2	3	4	5	6	7	8	9
				neutral				definitely hypothesis	

definitely <u>not</u> evidence	1	2	3	4	5	6	7	8	9
				neutral				definitely evidence	

why (explain): _____

completely disbelieve/reject	1	2	3	4	5	6	7	8	9
				neutral				completely believe/accept	

p) Other researchers think that dogs get an aggressive disorder because they lack a certain chemical.

definitely <u>not</u> hypothesis	1	2	3	4	5	6	7	8	9
				neutral				definitely hypothesis	

definitely <u>not</u> evidence	1	2	3	4	5	6	7	8	9
				neutral				definitely evidence	

why (explain): _____

completely disbelieve/reject	1	2	3	4	5	6	7	8	9
				neutral				completely believe/accept	

Problem 3

(Tell your interviewer you're on problem 3. Then read on.) We have a rule in mind that governs a set of three numbers, and we'd like you to try to guess what it is. At any time, you can either:

- Propose sets of three numbers, to each of which your experimenter will reply "that set of numbers fits the rule" or "that set doesn't fit the rule," or
- State what you think the rule is, if you think you've figured it out.

Here is a set of three numbers to start: **2,4,6**

Try to use what you know about disconfirmation!

(Give this sheet to your interviewer. He or she will write down your proposed numbers and rules in the order you propose them.)

Problem 4

Consider the following passage:

Some dogs have an aggressive disorder. They bark more than other dogs, growl at strangers, and sometimes even bite. They also tend to have higher blood pressure and heart rate than other dogs.

Some researchers think that these dogs get the aggressive disorder when their owners treat them poorly, that is, when the owner neglects the dog, doesn't give it enough love, or hits it. These researchers trained one group of aggressive-disorder dog owners to treat their dogs firmly yet lovingly. They found that all dogs whose owners were trained barked much less, were much friendlier to strangers, never bit a stranger, and had lower heart rate and blood pressure than dogs whose owners had

not been trained. These researchers said that their experiment proved that abuse causes dogs to have the disorder.

Other researchers disagree. They think that dogs with the disorder are born without a certain chemical in their body. They think that the lack of this chemical elevates their blood pressure and causes the disorder. These researchers gave one group of aggressive-disorder dogs a medicine that contained the chemical. They found that the dogs had a much lower heart rate and blood pressure, were friendlier to strangers, did not bark as much, and never bit anyone. These researchers said that their experiment proved that the missing chemical causes dogs to have the disorder.

With this passage in mind, and based your view and knowledge of the world, please rate the following statements as you did in Problem 2.

a) Some dogs have an aggressive disorder in which they bark more, growl more, bite more, and have higher blood pressure and heart rate than other dogs do.

definitely <u>not</u> hypothesis	1	2	3	4	5	6	7	8	9	definitely hypothesis
-------------------------------------	---	---	---	---	---	---	---	---	---	--------------------------

definitely <u>not</u> evidence	1	2	3	4	5	6	7	8	9	definitely evidence
-----------------------------------	---	---	---	---	---	---	---	---	---	------------------------

why (explain): _____

completely disbelieve/reject	1	2	3	4	5	6	7	8	9	completely believe/accept
---------------------------------	---	---	---	---	---	---	---	---	---	------------------------------

b) Some researchers think dogs get an aggressive disorder when their owners treat them poorly.

definitely <u>not</u> hypothesis	1	2	3	4	5	6	7	8	9	definitely hypothesis
-------------------------------------	---	---	---	---	---	---	---	---	---	--------------------------

definitely <u>not</u> evidence	1	2	3	4	5	6	7	8	9	definitely evidence
-----------------------------------	---	---	---	---	---	---	---	---	---	------------------------

why (explain): _____

completely disbelieve/reject	1	2	3	4	5	6	7	8	9	completely believe/accept
---------------------------------	---	---	---	---	---	---	---	---	---	------------------------------

c) Some researchers trained one group of aggressive-disorder dog owners to treat their dogs firmly yet lovingly.

definitely <u>not</u> hypothesis	1	2	3	4	5	6	7	8	9	definitely hypothesis
-------------------------------------	---	---	---	---	---	---	---	---	---	--------------------------

definitely <u>not</u> evidence	1	2	3	4	5	6	7	8	9	definitely evidence
-----------------------------------	---	---	---	---	---	---	---	---	---	------------------------

why (explain): _____

completely disbelieve/reject	1	2	3	4	5	6	7	8	9	completely believe/accept
---------------------------------	---	---	---	---	---	---	---	---	---	------------------------------

d) Some researchers found that training dog owners to treat their dogs firmly yet lovingly relieved symptoms of aggressive disorder in their dogs.

definitely <u>not</u> hypothesis	1	2	3	4	5	6	7	8	9	definitely hypothesis
-------------------------------------	---	---	---	---	---	---	---	---	---	--------------------------

definitely <u>not</u> evidence	1	2	3	4	5	6	7	8	9	definitely evidence
-----------------------------------	---	---	---	---	---	---	---	---	---	------------------------

why (explain): _____

completely disbelieve/reject	1	2	3	4	5	6	7	8	9	completely believe/accept
---------------------------------	---	---	---	---	---	---	---	---	---	------------------------------

e) Abuse causes an aggressive disorder in dogs.

definitely <u>not</u> hypothesis	1	2	3	4	5	6	7	8	9	definitely hypothesis
-------------------------------------	---	---	---	---	---	---	---	---	---	--------------------------

definitely <u>not</u> evidence	1	2	3	4	5	6	7	8	9	definitely evidence
-----------------------------------	---	---	---	---	---	---	---	---	---	------------------------

why (explain): _____

completely disbelieve/reject	1	2	3	4	5	6	7	8	9	completely believe/accept
---------------------------------	---	---	---	---	---	---	---	---	---	------------------------------

f) Other researchers think that dogs get an aggressive disorder because they lack a certain chemical.

definitely	<u>not</u>						definitely	
hypothesis	hypothesis	neutral					hypothesis	
1	2	3	4	5	6	7	8	9

why (explain): _____

completely	disbelieve/reject						completely	
1	2	3	4	5	6	7	8	9

g) Other researchers found that a chemical relieved symptoms of aggressive disorder in dogs.

definitely	<u>not</u>						definitely	
hypothesis	hypothesis	neutral					hypothesis	
1	2	3	4	5	6	7	8	9

why (explain): _____

completely	disbelieve/reject						completely	
1	2	3	4	5	6	7	8	9

h) Lack of a chemical causes an aggressive disorder in dogs.

definitely	<u>not</u>						definitely	
hypothesis	hypothesis	neutral					hypothesis	
1	2	3	4	5	6	7	8	9

why (explain): _____

completely	disbelieve/reject						completely	
1	2	3	4	5	6	7	8	9

List any other plausible hypotheses about what might cause the disorder (besides what the researchers thought caused it):

How might you test the hypotheses you wrote down?

Do the researchers' results prove anything? If so, what?

How much do you believe the following:

a) The experiment by the first set of researchers proved that abuse causes the aggressive disorder.

completely disbelieve/reject	neutral	completely believe/accept
1 2 3	4 5	6 7 8 9

b) The experiment by the second set of researchers proved that lack of the chemical causes the aggressive disorder.

completely disbelieve/reject	neutral	completely believe/accept
1 2 3	4 5	6 7 8 9

Problem 5

Using what you know about disconfirmation, consider the following situation:

You have four sheets of paper in front of you. Every sheet has a letter on one side and a number on the other side. Your task is to decide which sheets you need to turn over to determine the truth or falsity of this rule: *If there is a vowel on one side of a card, then there is an even number on the other side.*



What is the minimum number of cards you would need to turn over? What are they, if any?

Why did you choose the cards that you did?

Problem 6

Consider the following passage:

Chris and Pat are trying to name a wine. They both note the wine's sparkle, pink color, fruity taste, and strong aroma. Chris says that the Grapes of Frath wine is a common wine, so the wine is probably a Grapes of Frath wine. Chris also notes that if one assumes that it is a Grapes of Frath wine, that explains why it has sparkle and why it is pink in color.

Pat disagrees with Chris. Pat believes that it is a Pana Valley wine. Pat says that he thinks that their host likes Pana Valley wines, so the wine is probably a Pana Valley wine. Pat also notes that if one assumes that it is a Pana Valley wine, that explains why it is pink in color, why it has a fruity taste, and why it has a strong aroma.

a) List the hypotheses mentioned in the text. Label them H1, H2, H3, etc.....:

b) List the evidence mentioned in the text. Label them E1, E2, E3, etc.....:

c) Below, please rate how strongly you believe the statements you listed in (a) and (b), on a scale from 1 (completely disbelieve/reject) to 9 (completely believe/accept).

Write the label of the statement below, followed by the rating (e.g., H12: 3, E13: 5, etc...)

d) List (and label) any other plausible hypotheses, not mentioned in the text, that come to mind:

e) List (and label) any other relevant evidence, not mentioned in the text, that comes to mind:

f) Below, please rate how strongly you believe the statements you listed in (d) and (e), on a scale from 1 (completely disbelieve/reject) to 9 (completely believe/accept). Write the label of the statement below, followed by the rating (e.g., H12: 3, E13: 5, etc...)

g) What statements so far (in a, b, d, and e) *explain* what other statements? (You don't have to write the statements out. You can use the labels to say things like, "H5 explains E10")

h) What statements so far (in a, b, d, and e) *contradict* what other statements? (You don't have to write the statements out. You can use the labels to say things like, "H1 contradicts H4")

i) In the space below, please make any revisions to your argument that seem appropriate.

Add explanation(s):

Delete explanation(s):

Add contradiction(s):

Delete contradiction(s):

j) Revise your ratings for any of the statements (in a, b, d, and e) if that seems appropriate. Write your new ratings below (again, write the label of the statement, followed by it's rating).

Problem 7

Consider the following passage:

A boy wants to ask a girl to see a movie with him. Will Emily say yes or no to Zachary?

On one hand, Zachary believes that Emily may dislike him. Emily laughed at him when he fell on the baseball field. Emily did not talk to Zachary when he saw her at the mall. And when Zachary ran for class vice-president, Emily supported Zachary's opponent. Finally, the assumption girls are more prone to dislike boys than to like them suggests that she might not like Zachary. The possibility that Emily dislikes Zachary would mean she will say no to seeing a movie with him.

On the other hand, Zachary believes that Emily might indeed like him. Emily attends his baseball practices frequently. Sometimes Zachary catches her watching him in class. And he got a valentine from Emily in February. The possibility that Emily likes Zachary means that she will say yes to seeing a movie with him.

a) List the hypotheses mentioned in the text. Label them H1, H2, H3, etc.....:

b) List the evidence mentioned in the text. Label them E1, E2, E3, etc.....:

c) Below, please rate how strongly you believe the statements you listed in (a) and (b), on a scale from 1 (completely disbelieve/reject) to 9 (completely believe/accept). Write the label of the statement below, followed by the rating (e.g., H12: 3, E13: 5, etc...)

d) List (and label) any other plausible hypotheses, not mentioned in the text, that comes to mind:

e) List (and label) any other relevant evidence, not mentioned in the text, that comes to mind:

f) Below, please rate how strongly you believe the statements you listed in (d) and (e), on a scale from 1 (completely disbelieve/reject) to 9 (completely believe/accept). Write the label of the statement below, followed by the rating (e.g., H12: 3, E13: 5, etc...)

g) What statements so far (in a, b, d, and e) *explain* what other statements? (You don't have to write the statements out. You can use the labels to say things like, "H5 explains E10")

h) What statements so far (in a, b, d, and e) *contradict* what other statements? (You don't have to write the statements out. You can use the labels to say things like, "H1 contradicts H4")

i) In the space below, please make any revisions to your argument that seem appropriate.

Add explanation(s):

Delete explanation(s):

Add contradiction(s):

Delete contradiction(s):

j) Revise your ratings for any of the statements (in a, b, d, and e) if that seems appropriate. Write your new ratings below (again, write the label of the statement, followed by it's rating).

APPENDIX B: Unit 1, "Evidence, Hypotheses, and Theories"

Unit 1: Evidence, hypotheses, and theories

In this section, we'll talk about distinctions between **evidence**, **hypotheses**, and **theories**. Depending upon who you ask, the definitions for these concepts can vary greatly. From this section you should be able to get a good understanding of what each concept is even if you may not be able to describe it perfectly in words.

In general, we will define a **hypothesis** as being a statement that attempts to state what might be true in a particular situation. The word "attempts" in the previous sentence is important because some hypotheses accurately describe the segment of reality that they attempt to describe, while other hypotheses can distort, falsify, or misrepresent reality. In other words, some hypotheses are going to be correct while others are not. Here are some examples of hypotheses to help you out:

I will get wet if I don't have my umbrella.

Latrell Sprewell of the Golden State Warriors should be Rookie of the Year.

Recycling helps the environment.

Hypotheses are the essential units, or building blocks, out of which more complicated descriptions are made. This leads us to the discussion of **theories**. Theories are collections of hypotheses and **evidence** (which we will discuss soon). An example of such a theory is the Theory of Evolution. This theory consists of many hypotheses such as:

New species are formed by natural selection.

There is a progressive change, over time, in organisms from simple to complex.

Genetic traits are passed from parent to offspring.

Evidence is an extremely important part of any successful theory. Evidence is any body of factual statements on which a belief, idea, or hypothesis is based. Thus, the believability of a hypothesis depends upon the evidence and other hypotheses that support or explain it.

We may believe a hypothesis because it agrees with our own observations (evidence), or because other people have reported that they have made observations that agree with, or support, the hypothesis. All evidence is not created equal, though. For instance, the observation that the world appears to be flat (and not spherical) from our viewpoint is usually considered to be weaker evidence than the evidence that we obtain from photographs taken by orbiting satellites.

How good a piece of evidence is partly depends on how the evidence was obtained: Was a reliable method used? How good are the tools that were used to gather or measure the evidence? Can the data be replicated in another experiment? Is the evidence an acknowledged fact or statistic? Is it a reliable memory or observation? Might some reasonable people disagree with the evidence?

It is important to note that, all other things being equal, evidence should usually be believed over a contradictory hypothesis—unless the evidence is dubious or very unreliable. For example, if a solitary hypothesis contradicts a solitary piece of evidence, then that contradicting evidence should be more believable than the hypothesis. However, if a hypothesis has a lot of supporting evidence and one solitary piece of contradicting evidence, then the hypothesis might be considered to have more truth than the evidence.

For our example of the **Theory of Evolution** discussed earlier, examples of evidence are:

- fossil records
- sickle-cell anemia data (a hereditary disease)

- that human, bird, and reptile embryos share similar features

Exercise 1

What are the hypotheses (and possible hypotheses) in the paragraph below, and what is the evidence for them? Label the hypotheses and evidence for easy reference; for example {H1, H2, etc.} and {E1, E2, etc.}. Time doesn't matter, so relax and be creative. You can either label them in the text itself, or write them down on the next page.

Mike is an average student (C average). He is taking a physics class, in which he has turned in the homeworks, which are usually about half correct. His best friend, John, is also in the class. John is an excellent student, who is very diligent with his studies. John speaks out in class, but Mike does not. Both John and Mike received A's on their midterms. The teacher thinks that Mike may have cheated on the exam, since Mike was sitting next to John.

Hypotheses:

Evidence:

Relationships between hypotheses and evidence

In the last part, we talked about differences between hypotheses and evidence. It is possible for hypotheses to explain other hypotheses, as well as to explain evidence. Sometimes two or more hypotheses together (but not individually) explain a piece of evidence. For example, music coming from a room might be explained by Jeff singing (along), or by Jenny singing (alone). However, a duet cannot be explained by Jeff or Jenny singing alone, but only by them both singing. This sort of explanation is termed a **joint explanation**. Similarly, hypotheses and/or evidence can contradict each other. Remember that it is possible for one proposition to explain and/or contradict more than one other proposition. (A "proposition" or "belief" can be a hypothesis or a piece of evidence.)

The following is an exercise to let you practice determining the relationships between hypotheses and evidence. Remember the following short passage from the last section:

Mike is an average student (C average). He is taking a physics class, in which he has turned in the homeworks, which are usually about half correct. His best friend John is also in the class. John is an excellent student, who is very diligent with his studies. John speaks out in class, but Mike does not. Both John and Mike received As on their midterms. The teacher thinks that Mike may have cheated on the exam, since Mike was sitting next to John.

Exercise 2a

On the next page, draw a diagram representing the relationships between the (evidential and hypothetical) propositions you labeled in Exercise 1. Use a solid line to connect a proposition that explains another, a dashed line for beliefs that are in conflict, and converging lines for beliefs that jointly explain another proposition. (Glance at the example diagram three pages ahead if you are confused.)

Your diagram:

From your diagram, what seems to be the most believable, coherent theory (or set of beliefs)? Why?

Exercise 2b

Are there other factors, not present in the text, that might need to be taken into account? If so, what are they, and how would they affect your previous reasoning?

Add these new factors to your diagram, in a different color pen.

Exercise 2c

Based on this new diagram, what are the most reasonable conclusions you can form?

An example analysis

The following pages include one possible analysis of the above situation. Compare your analysis with the following, and be sure that you understand the differences. If you feel that your analysis needs refining, go ahead. But, remember that your analysis does NOT need to be the same as the one below.

Tanya lists some possible hypotheses and evidence for them:

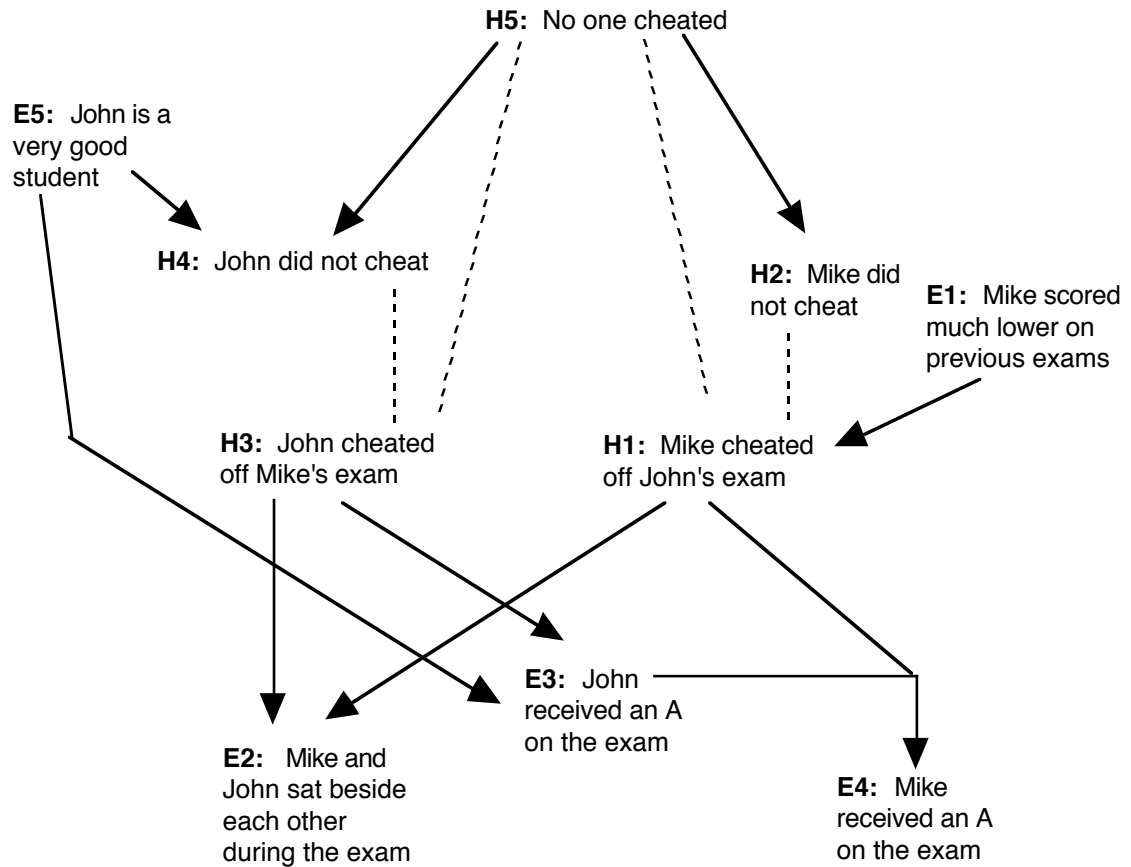
H1: Mike cheated off of John's exam.
 H2: Mike did not cheat.
 H3: John cheated off of Mike's exam.
 H4: John did not cheat.
 H5: No one cheated on the exam.
 E1: Mike scored much lower on previous exams.
 E2: Mike and John sat beside each other.
 E3: John received an A on the exam.
 E4: Mike received an A on the exam.
 E5: John is a very good student.

Here is Tanya's set of explanatory and contradictory relationships between the hypotheses and evidence:

H1 and E3 explain E4
 E1 explains H1
 H1 explains E2
 H3 explains E2
 H3 explains E3
 H5 explains H2
 H5 explains H4
 E5 explains H4
 E5 explains E3
 H1 contradicts H2
 H1 contradicts H5
 H3 contradicts H4
 H3 contradicts H5

Here is Tanya's diagram of the relationships between the hypotheses and evidence. In our convention, we have the things that do the explaining above the things they explain. You'll notice that we put arrows in the direction of the explanation. You might want to go back and add arrows to your diagram (and any

future diagrams), too. You don't have to write out your text in the diagram, like we do here. You can just use your labels. We're just writing out the text here for your ease.



Tanya thinks that no single hypothesis has overwhelming support:

"Both H1 and H3 seem possible. It is difficult to rule out H5, because of the notion of 'innocent until proven guilty.'"

Tanya tries to think of other ideas that might be important for understanding what happened. She came up with the following:

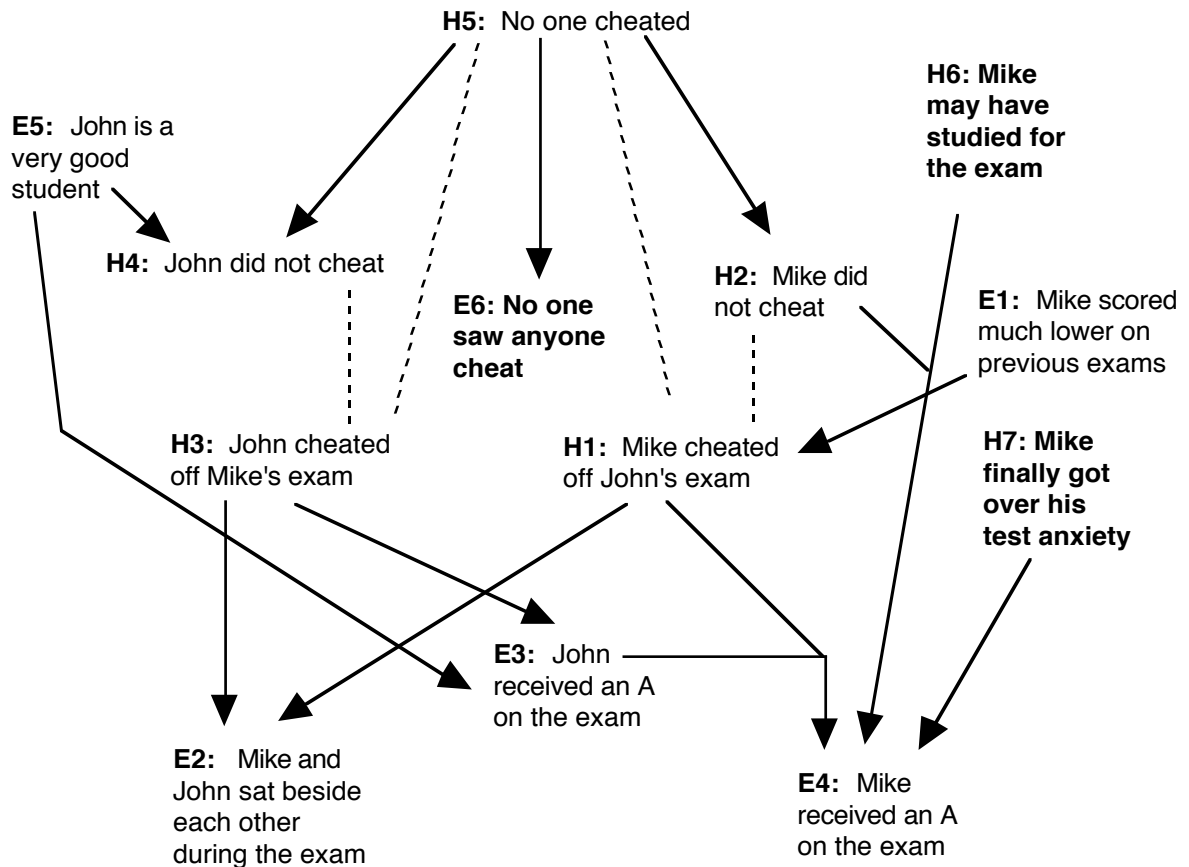
H6: Mike may have studied for the exam.

H7: Mike may have finally gotten over his test anxiety.
 E6: No one saw anyone cheat.

Here are Tanya's additional explanatory and contradictory relationships between the hypotheses and evidence:

H5 explains E6
 H6 and H2 explain H2
 H7 explains E4

And here is Tanya's new diagram, with new beliefs in boldface:



Based on this new diagram, Tanya may conclude:

"H1, H3, and H5 all have support, but the evidence is still not conclusive. It's not clear that we have an argument strong enough to accuse anyone of cheating."

Exercise 3

Consider the following passage, list the hypotheses and evidence. What hypotheses could be inferred from this new information? Add them (along with any explanatory and contradictory relationships), with a different color pen, to your diagram.

John's girlfriend, Mary, is also in the physics class. Mary is a solid B student, but she did very poorly on the exam. Mary and John had studied for the midterm together and were sitting next to each other during the exam.

APPENDIX C: Unit 2, "Reasoning About Arguments"

Unit 2: Reasoning About Arguments

In this section, we'll talk more about the need for alternative hypotheses, generating arguments based on a given (scientific or everyday) controversy, and common "biases" in reasoning.

The need for alternative hypotheses

It's not always easy to come up with more than one hypothesis in a situation. People often get "stuck" and can only come up with one hypothesis (and many times they can't come up with *any* plausible hypotheses!) Nevertheless, most people would agree that it is good to have more than one alternative to consider in a given situation. Why? For one reason, considering multiple hypotheses can help you overcome **confirmation bias**. Confirmation bias is when you focus on trying to prove that your favorite hypothesis (or set of hypotheses) is correct, and don't try to see what could be wrong with it (or them). In other words, it's when you don't try to critique or **disconfirm** your hypotheses.

Why is it important to disconfirm your hypotheses? Because no matter how much evidence you see that agrees with your hypothesis, there are other hypotheses that could agree with the same evidence. So if you find a lot of evidence that supports a certain hypothesis, that's great and the hypothesis might very well be true. But, some other hypothesis might be true instead. Let's look at an example:

Four people who ate in the school cafeteria got sick. Patricia thinks it must have been the meatloaf that made them sick. She asks each of the four whether or not they ate meatloaf. They all said "yes," so Patricia decides she must have been right. She tells the principal that everyone who ate the meatloaf will have to go to the hospital.

Here is a case in which Patricia has a hypothesis that is consistent with the data. She may be right—it may very well be the meatloaf. But can we be *sure*? If you were the principal, would you send everyone who ate the meatloaf to the hospital? How would you decide if Patricia were right? Consider the following:

Christopher doesn't think the meatloaf was the problem. He suspects the mashed potatoes were the problem. So he asks the four people who got sick whether they had mashed potatoes. They all say yes. So Christopher goes to the principal and tells her that everyone who ate mashed potatoes should be sent to the hospital.

The problem that the principal faces is that there are (at least) two hypotheses that are consistent with the evidence. What can the principal do to learn more about what might have caused the illness? The principal decides to try to disconfirm the hypotheses:

The principal asks some other students, who ate at the same time but didn't get sick, if they had eaten the meatloaf or potatoes. She finds that many other people ate both the meatloaf and mashed potatoes but didn't get sick!

The principal found out that both Patricia and Christopher were wrong, by trying to disconfirm their hypotheses (and succeeding!). What do scientists do if all the hypotheses are disconfirmed? They think of more possibilities! The tough part is that there are always more hypotheses to think up. So scientists try to think up as many plausible, reasonable hypotheses as they can *early on*. That makes it more likely that, when they seek disconfirmation and eliminate hypotheses, the correct one will be left standing:

The principal decides that before she sends anyone to the hospital, she should think up as many possibilities as she can for what caused the students to get sick. The illness could have come from the butter, the ketchup, the chocolate pudding, or anything else the four sick students ate. Then

she can check to see if other people got sick from any of those foods.

How do you decide if you have enough hypotheses? In our case, the principal did think up lots of alternative hypotheses. But is the correct one there? Perhaps the students are ill because of something they ate at the football game the night before, or maybe even because they are all friends with the same person and caught a flu from him. There is no easy way to know when you have enough hypotheses to investigate. This is one reason that many scientists study the same questions. You have to use your best guesses about what is likely in order to choose which hypotheses to look into. The principal didn't, for instance, investigate space aliens as a possible reason that the four students were sick—nor would we want her to!

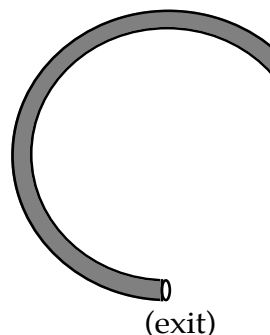
The moral of the story is: You can only be sure something is wrong, never that something is completely right. Another hypothesis might be the true one, and your best guess might be false or incomplete—you just haven't seen enough evidence to be able to know for sure. That's why it's best to come up with as many hypotheses as you can right in the beginning, then work hard to disprove each of them. If you run out of hypotheses because you've disproved all of the ones you have, it's back to the drawing board to come up with new possible explanations. However, if exactly one hypothesis or theory remains standing and all its competitors are disproven, then it's likely that hypothesis or theory is close to the truth.

Exercise 1

Consider the following situation:

A ball is rolling quickly, counter-clockwise, through a tube that is lying flat on a table. The tube is bent into a C shape, as shown.

(Top view)



Draw as many plausible, alternative paths that you or someone else might think the ball could shoot out.

Generating arguments based on a scientific or everyday controversy

One of the first steps to solving any problem is gathering all the relevant information. The same applies to considering scientific problems. Remember back to Unit 1 where you diagrammed the following situation:

Mike is an average student (C average). He is taking a physics class, in which he has turned in the homeworks, which are usually about half correct. His best friend John is also in the class. John is an excellent student, who is very diligent with his studies. John speaks out in class, but Mike does not. Both John and Mike received As on their midterms. The teacher thinks that Mike may have cheated on the exam, since Mike was sitting next to John.

You diagrammed this problem by writing down the hypotheses and evidence that you were told about in the text, for instance that "Mike is usually half correct." or that "John is an excellent student" or hypotheses like "Mike cheated off John on the

exam" or "Mike was honest on the exam." Some of those ideas were not actually mentioned in the text: for instance, the hypothesis that "Mike was honest on the exam" was not specifically mentioned anywhere in the text. But one could infer it, since it is the obvious alternative to the teacher's idea that Mike cheated.

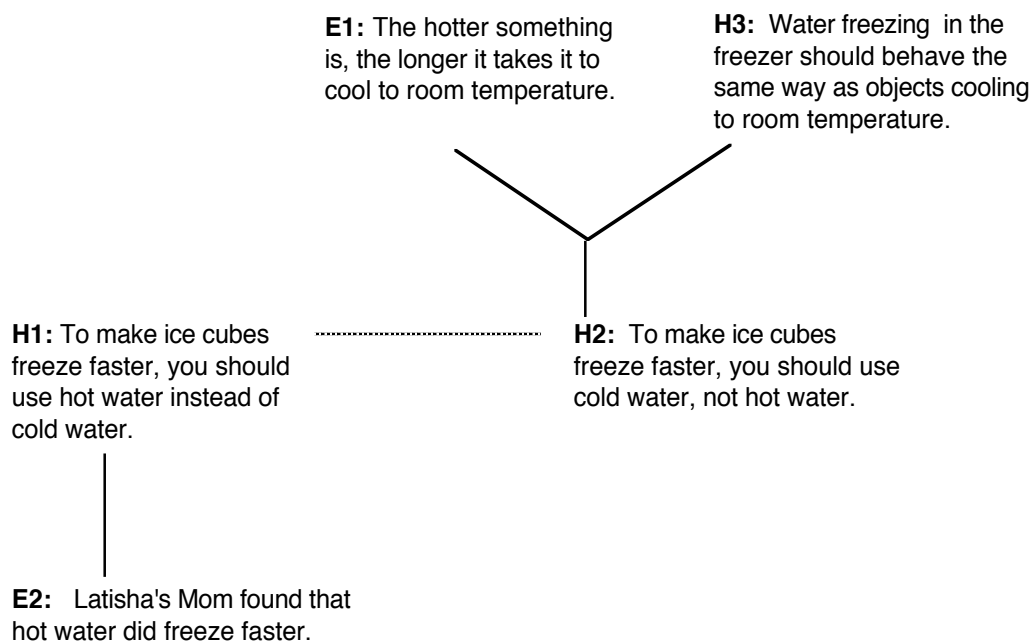
Everyday situations don't come with texts that describe all the relevant information. Scientific situations don't, either. Often, the amount that is implicit or implied is much larger than the amount that is explicitly discussed. Learning to notice which things are implied is a very important part of understanding a situation. Here are two different viewpoints about a common situation: freezing ice cubes.

Latisha's Mom says that to make ice cubes freeze faster, you should use hot water instead of cold water in the ice cube tray. She has been doing this for many years, and although she didn't believe it when she first heard it, Latisha's Mom tried it out several times and the hot water did freeze faster.

Latisha learned in science class that it takes longer for hot things to cool to room temperature than it takes for warm things which are closer to room temperature.

How would you diagram this situation? There are a lot of important ideas that are left unmentioned. For instance, what might Latisha's hypothesis be, regarding ice cubes? Why would she think that? How does what she knows about objects cooling to room temperature tell her anything about freezing water in the freezer? You would have to say what hypotheses, everyday evidence, and common facts (another kind of evidence, really) Latisha and her Mom might be using to think about the situation.

Here is such a possible diagram:



Notice how Latisha used her knowledge about a similar situation and her belief that water in the freezer should behave the same way. This is a simple **analogy**, and analogies are often quite useful.

As you can see, Latisha's hypothesis H2 has a good point that is jointly explained by evidence E1 and hypothesis H3. But she doesn't have any direct evidence about ice cubes. Latisha's Mom does have some direct evidence (E2). But she doesn't have an explanation for her hypothesis H1. Latisha has a really hard time believing her Mom. Latisha could do a number of things in this situation. She could try it out for herself. She could look in books or ask her science teacher if there were some reason objects cooling to room temperature would be different than the freezing ice cubes. Instead, she does something many scientists do: She questions how good her Mom's evidence is.

Latisha adds another idea to her graph of the argument. "Mom doesn't measure carefully."

How do you feel about this argument—is Latisha or her Mom right? What would it take to convince you of the other viewpoint?

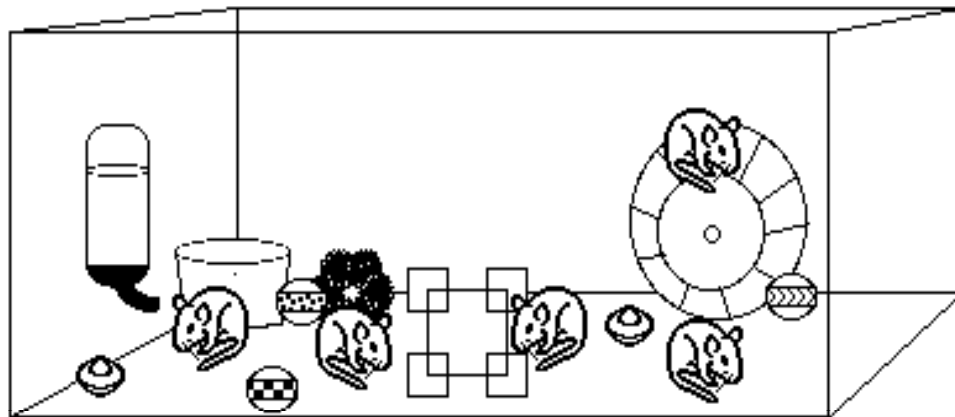
As it turns out, hot water *does* freeze faster than cold water. One explanation is that, since more of the hot water evaporates, there is less water to freeze; so it takes less time to freeze than the (eventually more massive) cold water. But that's just one hypothesis. Perhaps you can think of alternate hypotheses that explain how the hot water actually freezes faster, and how you might test these hypotheses. In any case, having an explanation for why a surprising piece of evidence is true makes a big difference in how easy it is to believe the evidence. For example, having an explanation of why the hot water freezes faster might make Latisha not question her Mom's measurements. Without a good explanation, she might have just rejected her Mom's experience.

Exercise 2

Consider the following passage:

A UC Berkeley researcher believed that interesting, educational experiences in early life lead to larger brains. She found that rats raised alone in the empty cages had smaller brains than the rats raised together in the interesting environment. Based on this experiment, she concluded that children who have interesting, educational experiences in preschools will grow up to be more intelligent adults than children who do not attend preschool.

A preschool teacher disagreed with the researcher. She said that the rat experiment could not be used to explain the advantages of attending preschool.

**Exercise 2a**

List the hypotheses mentioned in the text. Label them H1, H2, H3, etc.....:

List the evidence mentioned in the text. Label them E1, E2, E3, etc.....:

Exercise 2b

Rate how strongly you believe each of the statements that you wrote above, on a scale from 1 (completely disbelieve/reject) to 9 (completely believe/accept). (You can write the ratings to the right of the statements.)

Exercise 2c

List (and label) any other plausible hypotheses, not mentioned in the text.

List (and label) any other evidence that comes to mind.

Exercise 2d

What statements *explain* what other statements? (You don't have to write the statements out. You can use the labels to say things like, "H5 explains E10")

What statements *contradict* what other statements? (You don't have to write the statements out. You can use the labels to say things like, "H1 contradicts H4")

Exercise 2e

Revise your ratings if you want, by writing your new rating **to the right** of your old rating. (Don't scratch your old ratings out!)

Common biases in reasoning

Earlier we mentioned **confirmation bias**, which happens when you focus only on trying to confirm your favorite hypotheses and don't try to disconfirm them. In the cafeteria example above, Patricia and Christopher showed some confirmation bias, because unlike the principal, they didn't ask questions of students who *didn't* get sick. It's not irrational to try to confirm a hypothesis—when scientists come up with a new hypothesis, sometimes they focus first on confirming it, and later they try to disconfirm it. But *neglecting* disconfirmation overall can lead you to accept a hypothesis that might be false or incomplete, as we showed in the cafeteria example. (Since confirmation bias is common, we don't need to worry as much about neglecting confirmation!) The moral of the story is: try to seek a *balance* between confirmation and disconfirmation.

Two other common biases have to do with the *order* in which hypotheses and evidence are considered or gathered. For example, sometimes people tend to cling to previous information (e.g., their original beliefs) and ignore or discount new information. This is sometimes called a **primacy bias** (or **intransigence**). Being cautiously objective and thinking up *plausible alternate hypotheses* usually helps to reduce primacy bias, as does being *objective* about the alternatives and trying to confirm *and* disconfirm all plausible hypotheses. Consider the follow example:

Before the principal disconfirmed Patricia's and Christopher's hypotheses about what was making the students sick, Christopher went to Patricia and told her that he thought the mashed potatoes, not the meatloaf, were the problem. He told her that he asked four of the sick students whether they had mashed potatoes, and they all said yes. Patricia thought about it a while, and said that her hypothesis could be wrong, but she still thought that it was better than Christopher's. She said that maybe the meatloaf was still the problem, and maybe the people Christopher talked to ate the meatloaf, too.

Here is a case where an alternative hypothesis makes Patricia lose some faith in her initial hypothesis, but she still doesn't want to recognize Christopher's

hypothesis as an equally plausible alternative. How would you convince Patricia that at this point, Christopher's hypothesis is *just as plausible* as hers?

Christopher told Patricia that he thought that his hypothesis was just as plausible as hers, since the people that Patricia talked to might just as well have eaten the mashed potatoes too. He said that maybe she thought her hypothesis was better just because she considered that one first and wanted to be right, but both hypotheses were equally plausible.

Christopher is trying to get Patricia to be more *objective* about the two hypotheses. Viewing hypotheses merely as "objects of reasoning," that is, trying to not favor beliefs in which you have a vested interest, can help to reduce primacy bias. While it might be nice if your pet beliefs are "right," in the long run it's usually more interesting and useful to try to figure out which hypotheses best explain the situation.

The principal told Patricia that several students who didn't get sick had eaten the meatloaf and mashed potatoes. Patricia found Christopher and told him what the principal said, and that she now agreed that both of their hypotheses had been equally plausible, even though they were both wrong.

Here we see that, as a result of objectively considering and testing both hypotheses (and disconfirming both), Patricia comes to see that both hypotheses were equally plausible. Exercising caution, thinking up and objectively considering plausible alternate hypotheses, and trying to confirm and disconfirm the hypotheses will also probably help *you* reduce primacy (and confirmation) bias, and increase your chances of eliminating highly unlikely hypotheses.

At the other extreme, **recency bias** is the tendency to be more swayed by recent information and to discount previous findings. For example, the previous example could have gone this way instead:

Before the principal disconfirmed Patricia's and Christopher's hypotheses about what was making the students sick, Patricia

went to Christopher and told him that she thought the meatloaf, not the mashed potatoes, was the problem. She told him that she asked four of the sick students whether they had meatloaf, and they all said yes. Christopher thought about it a while, and said that her hypothesis was probably better than his.

Here is a case where Christopher too quickly accepts an alternative hypothesis. How could Patricia convince Christopher at this point that his hypothesis is just as plausible as hers?

Patricia told Christopher that she thought his hypothesis was just as plausible as hers. She had asked four people, and he had asked four people and both hypotheses had pretty much the same amount of good evidence. Patricia also told Christopher that since neither of them have any counter-examples, like someone who's fine who had meatloaf and/or mashed potatoes, they are probably equally likely to be wrong, but you can't tell. (The principal later showed this with counter-examples.) Christopher said that he saw her point, and that they might even *both* be right; it could be some strange combination of meatloaf and potatoes!

Here we see that laying out the plausible alternatives *explicitly* all at once, and cautiously and objectively considering these alternatives, can help reduce recency bias (as well as the other biases we've mentioned).

Summary

Sometimes when you try to confirm or disconfirm a hypothesis, what you discover may lead you to change your mind about things. You might decide on one or more of the following things:

- your hypothesis is wrong,
- your hypothesis is incomplete,
- your hypothesis might still be right if you revise it a little,

- some evidence that conflicts with or supports the hypothesis might not be very good, maybe because the method used to gather the evidence wasn't very reliable, and/or
- the structure of the entire argument, of which the hypothesis is a part, should be changed.

But it's not always clear how to form or change your opinion about a hypothesis or piece of evidence, especially when there are several plausible hypotheses and the situation or argument is complex. Here are some strategies you can use to help yourself:

- have an open mind and come up with as many hypotheses as you can,
- be ready to revise your hypotheses,
- work with other people to get their ideas and understand their hypotheses,
- write down your arguments explicitly so you can look at them for gaps, and/or
- remember to try to disconfirm each hypothesis—if one is basically correct, it will be able to withstand your best efforts. The truth "takes a lickin', but keeps on tickin'!"

Exercise 3

Consider the following historical example:

In 1915, Alfred Wegener proposed his theory of "continental drift." He claimed that continents slowly drift over the earth's surface, fracturing and re-uniting. He said that this drifting would explain the migration of some mammals, and that the forces generated from continental collisions could explain how mountains developed.

Wegener's contemporaries disagreed with him, claiming that each continent's location is fixed. They argued that the earth, once hot, has been cooling and contracting, and that the compressive forces generated from these contractions could have created mountains. Also, they said that Wegener did not offer a compelling explanation of what force(s) would cause the continents to move, or drift, like he proposed. Most scientists held this "rigid earth" theory until the 1960's, but now Wegener's theory is the established view.

Exercise 3a

List the hypotheses mentioned in the text. Label them H1, H2, H3, etc.....:

List the evidence mentioned in the text. Label them E1, E2, E3, etc.....:

Rate how strongly you believe each of the statements that you wrote above, on a scale from 1 (completely disbelieve/reject) to 9 (completely believe/accept).

Exercise 3b

List (and label) any other plausible hypotheses, not mentioned in the text.

List (and label) any other evidence that comes to mind.

Exercise 3c

What statements *explain* what other statements?

What statements *contradict* what other statements?

Exercise 3c

Revise your ratings if you want. Write your new rating **to the right** of your old.

Exercise 3f

What bias or kind of biases might explain why Wegener's peers discounted his theory? What else might Wegener have done to try to convince his peers?

APPENDIX D: Unit 3, "Using *Convince Me*"

Unit 3: Using **CONVINCE ME**⁸

What is **CONVINCE ME**?

CONVINCE ME is a computer program to help you think about your own reasoning. The program lets you type in short, sentence-like statements: things you believe and are sure of, and beliefs/things you're not so sure of. Then you can tell the computer which ideas explain and contradict the other ideas (see Figure 1).

So what? Why do I need a computer for that?

You don't. But just as explaining something to another person can help you understand something, entering an argument into **CONVINCE ME** can help you clarify your own beliefs. Also, just as people will often tell you what they agree and disagree with in your argument, **CONVINCE ME** will, in a similar way, tell you which statements your argument helps to affirm or reject and which ones it leaves neutral, from the computer's point of view.

How does the computer know what to believe?

It doesn't, except for what you tell it. When you put a statement in the computer, you'll be asked whether it is a piece of evidence or a hypothesis. Decide carefully, since the computer gives more weight to all pieces of evidence and then tries to figure out which hypotheses and evidence "hang together" best. **The computer doesn't understand the meanings of the statements that you type in.** It just tries to figure out which statements to believe on the basis of your argument—by what you tell it about what

⁸**CONVINCE ME** was developed by the **ECHO Educational Program (EEP)**, at the University of California, Berkeley. © 1993 University of California.

contradicts what, and what explains what. CONVINC ME uses a computer program called ECHO to do this.

What is "ECHO"?

ECHO is a computer model based on a theory called the "Theory of Explanatory Coherence" (TEC). The next section describes TEC and ECHO in more detail.

CM (old, no graphing)

Statements: Add... Edit... Delete Rate... Rate All... Model's fit...

Hypotheses:

Hypothesis	Rating	FCHO
H1. To make ice cubes freeze faster, use hot water, not cold water	5	6.7
H2. To make ice cubes freeze faster, use cold water, not hot water	6	4
H3. Water in the freezer should behave the same way as objects cooling	7	5.5

Evidence:

Evidence	Rating	FCHO
E1. The hotter something is, the longer it takes it to cool to room temperature	8	7.3
E2. Latisha's Mom found that hot water did freeze faster	7	7

Simulation results:

Hypotheses:

H1(6,7) H2(4) H3(5,5)

Evidence:

E1(7,3) E2(7)

Explanations: Explain... Explain All... Delete Explanation

The statement(s) that explain(s) **H2. To make ice cubes freeze faster, use cold water, not hot water** is/are:

H3. Water in the freezer should behave the same way as objects cooling to room temperature **AND**
 E1. The hotter something is, the longer it takes it to cool to room temperature

Contradictions: Conflict... Conflict All... Delete Conflict

The statement(s) that conflict(s) with **H2. To make ice cubes freeze faster, use cold water, not hot water** is/are:

H1. To make ice cubes freeze faster, use hot water, not cold water

Oops! (undo)

HelpMessages:

E1 The hotter something is, the longer it takes it to cool to room temperature

Current File:

Steps for using CONVINC ME:

1. Enter hypotheses and evidence.
2. Enter explanations and contradictions.
3. Rate the believability of your statements.
4. Run the ECHO simulation.
5. Compare your evaluations to ECHO's.
6. (optional) Make changes based on ECHO's feedback.

The correlation between your ratings and ECHO's evaluations is: 0.34 (mildly related).

The three most disparately rated statements are: H2, H1, H3, respectively (see boldened statements).

Your statement:

More of the hot water evaporates so there's less mass to freeze

Check all that apply:

- Acknowledged fact or statistic
- Observation or memory
- One possible inference, opinion, or view
- Some reasonable people might disagree

Select one:

Evidence E3 Reliability, if evidence? (from 1, poor, to 3, good)

Hypothesis H4

OK Cancel

Figure 1. The CONVINC ME program.

TEC and ECHO (you can skim the next two pages if you like)

The Theory of Explanatory Coherence (TEC) attempts to account for how people decide the plausibility of beliefs asserted in an explanation or argument. The theory is based on a few "hall of fame" **principles of reasoning**, such as:

- 1) The believability of an idea generally increases with increasing simplicity. In other words, making lots of (that is, joint) assumptions is often counterproductive, compared to making fewer assumptions.
- 2) People tend to believe statements when there is more evidence to support them.
- 3) We are more likely to believe something that doesn't conflict or compete with other things we strongly believe.

Etc. To learn more about TEC's principles, see the Appendix.

ECHO is a computer model based on TEC. In ECHO, arguments are represented as **networks of nodes** (like knots in a net). A hypothesis or piece of evidence is represented by a node, and explanatory or contradictory relations are represented by **links** between nodes. Hypothesis evaluation is treated as the satisfaction of **constraints** determined from the explanatory relations (that is, explanations and/or contradictions), TEC's principles, and from a few **numerical parameters**. Given a network of statements and relations between them, node activations are updated in parallel using a simple "**connectionist**" settling scheme. When the network of statements settles (or stabilizes), the nodes representing the most mutually coherent hypotheses and evidence are active, and the nodes representing inconsistent rivals are deactivated.

For example, suppose Chris says:

"Some people think that all animals (including humans) were created in their present form, about 5000 years ago. Others believe that animals evolved from earlier life slowly, over

millions of years. Both beliefs explain why animals exist. However, only the latter, evolutionary, hypothesis explains why transitions between forms in the fossil records appear to be gradual, and why scientists have found some fossils they estimate are over a million years old.

This could be represented in ECHO as:

hypothesis H1: "Animals were created in their present form about 5000 years ago."

hypothesis H2: "Animals evolved from earlier life over millions of years."

evidence E1: "Animals exist."

evidence E2: " Transitions between forms in the fossil records are gradual."

evidence E3: "Scientists have dated some fossils at over a million years old."

H1 competes with H2.

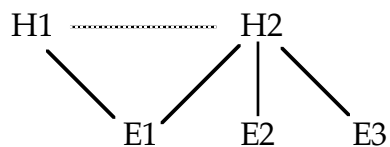
H1 explains E1.

H2 explains E1.

H2 explains E2.

H2 explains E3.

Or, in graphical network form (where solid lines represent explanatory links, and the dashed line represents a competing/contradictory link):



Given a scenario such as this, ECHO generates a numerical value for each statement that indicates how much it believes the statement. In general, the more positive the value, the more ECHO "believes" the statement; the more negative the value, the more ECHO "disbelieves" the statement. In this case, ECHO believes H2 over H1 since H2 explains more of the evidence.

**Please STOP skimming (and start reading thoroughly again)
from here on!**

Getting Help from CONVINC ME

If you have any questions about CONVINC ME or ECHO, select **About Convince Me...** or **About ECHO...** in the **Help** menu. To see a glossary of terms, select **Glossary...** in the **Help** menu. A glossary is also included at the end of this document. To see a summary list of steps about how to use CONVINC ME, select the **Steps...** item. When the steps are displayed, a checkmark (✓) will show up beside **Steps...** in the menu. This document will go through these steps in detail.

You can also use the **Help** menu to turn **Help Mode** on or off. When help mode is on, a checkmark (✓) will show up beside **Help Mode** in the menu, and messages will show up in the **Help/Messages** window (see Figure 2) when you pass the mouse cursor over parts of the software.

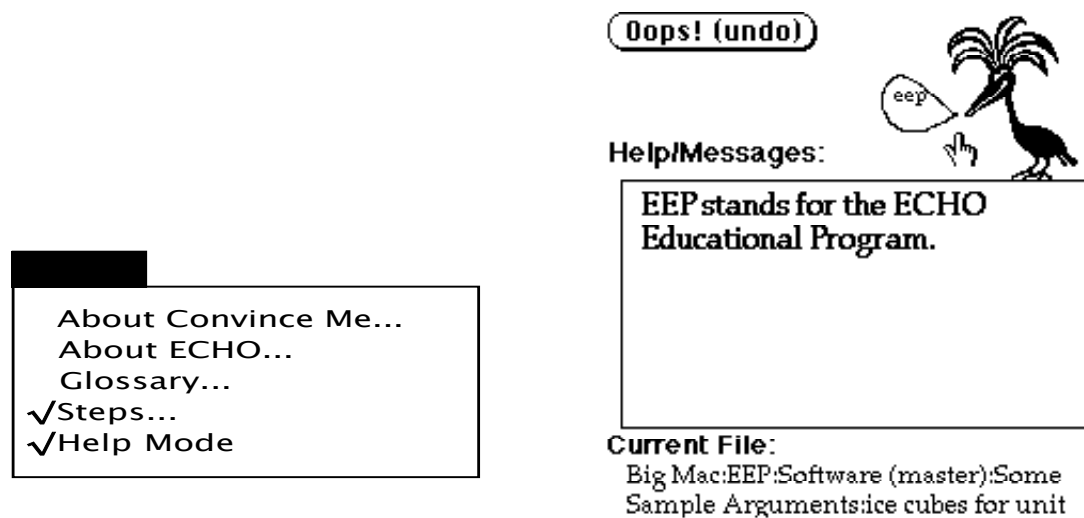


Figure 2. The Help menu and "Help/Messages window."

CONVINCE ME also remembers the things you do, and will "undo" the last thing that you did if you press the **Oops! (undo)** button. For example, if you delete a statement by mistake and want to bring it back, press **Oops! (undo)**. If you press **Oops! (undo)** a second time, it will "undo the undo"—that is, delete the statement (again).

Entering an argument

The Argument menu lets you create a new argument, load an existing argument, or save your argument (see Figure 3).

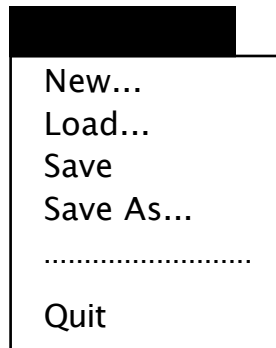


Figure 3. The Argument menu

Statements: Add... Edit... Delete Rate... Rate All... Model's fit...

Hypotheses: Rating ECHO

H1. To make ice cubes freeze faster, use hot water, not cold water	↑		↑		↑
H2. To make ice cubes freeze faster, use cold water, not hot water					
H3. Water in the freezer should behave the same way as objects cooling					
	↓		↓		↓

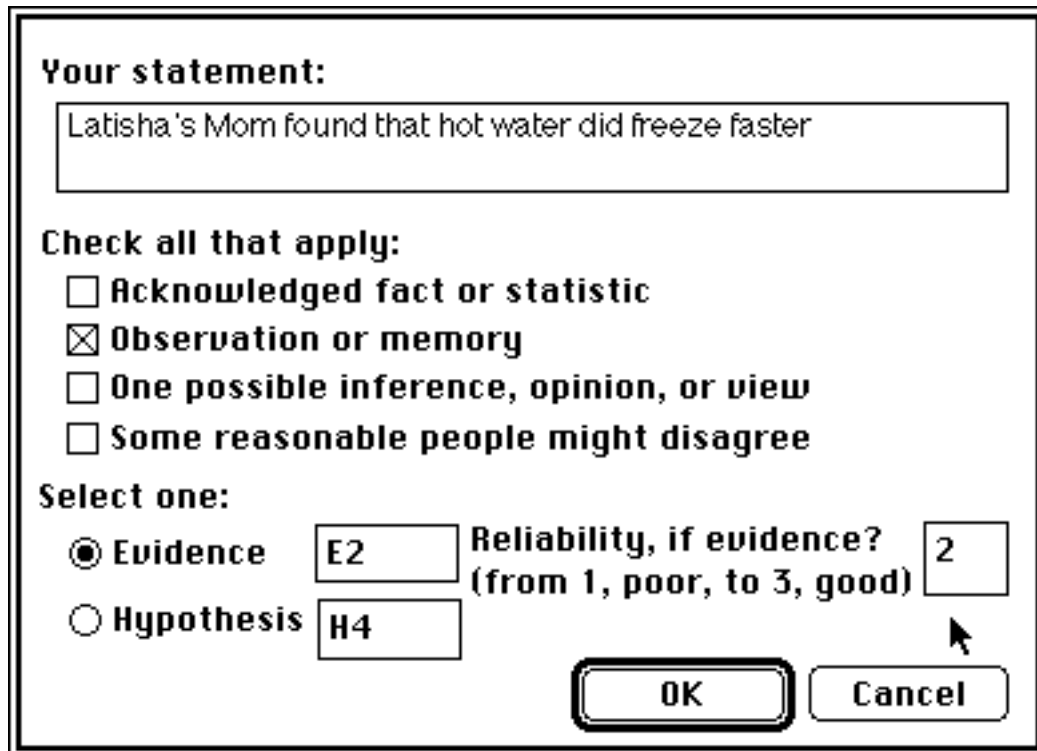
Evidence:

E1. The hotter something is, the longer it takes it to cool to room temper	↑		↑		↑
E2. Latisha's Mom found that hot water did freeze faster					
	↓		↓		↓

Figure 4. "Statements" window, with part of an argument we saw earlier.

When you want to enter statements for your argument, click the Add... button in the upper left section of the CONVINC ME screen (see Figure 4). A "dialog box" will then ask you what statement you would like to add (see Figure 5). It will also ask you to check one or more of the boxes to help determine if the statement is a hypothesis or a piece of evidence, and it also asks you to explicitly decide which one it is. (You may check no boxes if none really apply at all.) If the statement is a piece of evidence, CONVINC ME also wants to know how "reliable" you think the it is, on a scale from 1 (not very reliable) to 3 (very reliable):

Reliability (if evidence) is: poor fair good
 1 2 3



Your statement:

Latisha's Mom found that hot water did freeze faster

Check all that apply:

Acknowledged fact or statistic

Observation or memory

One possible inference, opinion, or view

Some reasonable people might disagree

Select one:

Evidence E2 Reliability, if evidence? (from 1, poor, to 3, good) 2

Hypothesis H4

OK Cancel

Figure 5. "Dialog box" to add or edit a statement.

If you want to change the text of a statement, or reclassify it as hypothesis or evidence or vice versa, click on the statement you want to modify and then click the **Edit...** button. If you want to delete a statement, select the statement and then click the **Delete** button (see Figure 4). After you've entered some statements, you can specify some explanations and contradictions among them.

Exercise 1

Create a new argument by selecting **New** from the **Argument** menu. Using the **Add...** button, add the following hypotheses and evidence to your argument (from the "ice cubes" argument in Unit 2). Don't specify any explanations and contradictions yet.

Hypotheses:

- To make ice cubes freeze faster, use hot water, not cold water (H1).
- To make ice cubes freeze faster, use cold water, not hot water (H2).
- Water in the freezer should behave the same way as objects cooling to room temperature (H3).

Evidence:

- The hotter something is, the longer it takes it to cool to room temperature (E1).
- Latisha's Mom found that hot water did freeze faster (E2).

Adding and deleting explanations

To create an explanation, you can select a statement in the "**Statements** window" that you want to explain (e.g., if you want to explain the statement, "To make ice cubes freeze faster, use cold water, not hot water," then click on it) and then click on the **Explain...** button in the "**Explanations** window" (see Figure 6). Alternately, you can also just click the

Explain All... button and let CONVINCE ME ask you for explanations for all of the statements, one after the other.

A "dialog box" will then come up with a list of statements, and ask you to specify your explanations (see Figure 7). You can select multiple statements by holding down the Command key when you click on a statement. (The Command key is the one that has the funny clover-leaf on it, between the option key and the spacebar.)

To delete an explanation, select it and then click the Delete Explain button in the "Explanations" window."

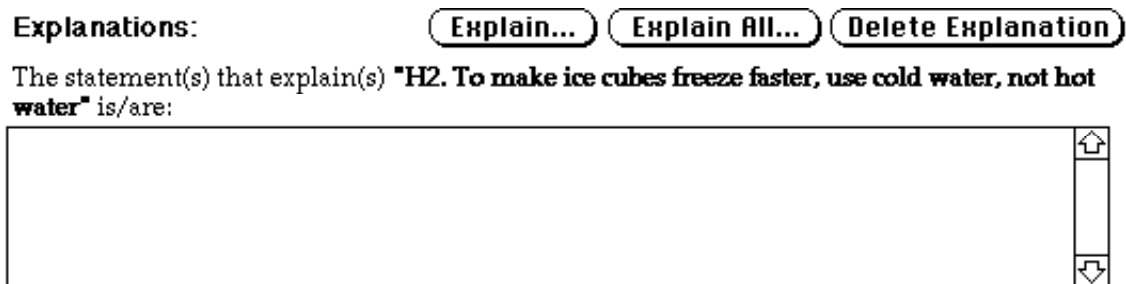


Figure 6. "Explanations" window before adding explanations.

What (if anything) explains the statement:
H2. To make ice cubes freeze faster, use cold water, not hot water

(Use command-click to select more than one statement.)

H1. To make ice cubes freeze faster, use hot water, not cold water
H3. Water in the freezer should behave the same way as objects cooling to room tem...
E1. The hotter something is, the longer it takes it to cool to room temperature
E2. Latisha's Mom found that hot water did freeze faster

Choose one:

Each statement explains the claim **independently**
 Statements **jointly** explain the claim

OK Cancel

Figure 7. "Dialog box" for adding explanations.

When you enter an explanation, the computer will ask you if the explanations that you select *independently* or *jointly* explain your claim.

If you click "**Each statement explains the claim **independently****", this means that each statement explains your claim *on its own*, i.e., "<statement one> explains the claim", and "<statement two> explains the claim.." etc. (E.g., That Todd was singing explains why music was coming from the room, and that Mary was singing also—independently—explains why music was coming from the room.)

Click "**Statements **jointly** explain the claim**" if the statements *together, in conjunction*, explain the claim" (that is, <statement one> alone doesn't explain the claim, but together with the other statement(s) you get a proper explanation; e.g., Todd singing and Mary singing jointly explains why it sounded like a duet).

Your explanations then appear in the explanations window (see Figure 8).

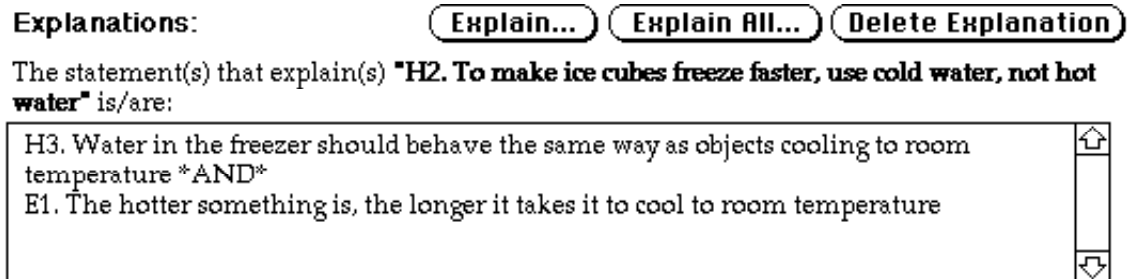


Figure 8. "Explanations" window after the first addition above.

Exercise 2

Add the following explanations to your ice cubes argument. To add the first explanation, click on H2 and then on the Explain... button, select E1 and H3 from the dialog box, and click on Statements *jointly* explain the claim. To add the second explanation, click on E2 and then on the Explain... button, select H1 from the dialog box, and click on Each statement explains the claim *independently*.

E1 and H3 jointly explain H2
H1 explains E2

Adding and deleting contradictions

To specify contradictions, you can select a statement in the "Statements window" that you want to contradict (e.g., if you want to specify what conflicts with the statement "To make ice cubes freeze faster, use cold water, not hot water," then click on it) and then click on the Conflict... button in the "Contradictions window" (see Figure 9). Alternately, you can also just click the Conflict All... button and let CONVINC ME ask you, for each statement, what conflicts with it.

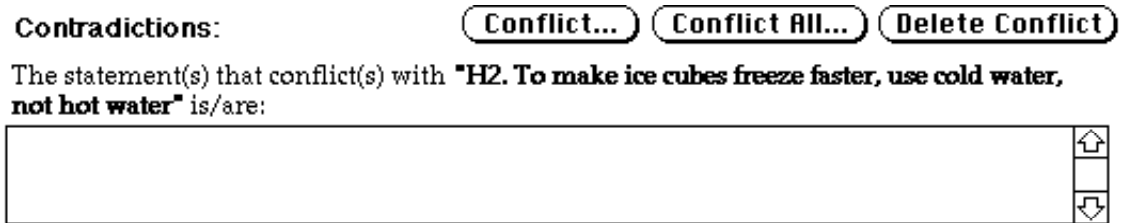


Figure 9. "Contradictions" window before adding contradictions.

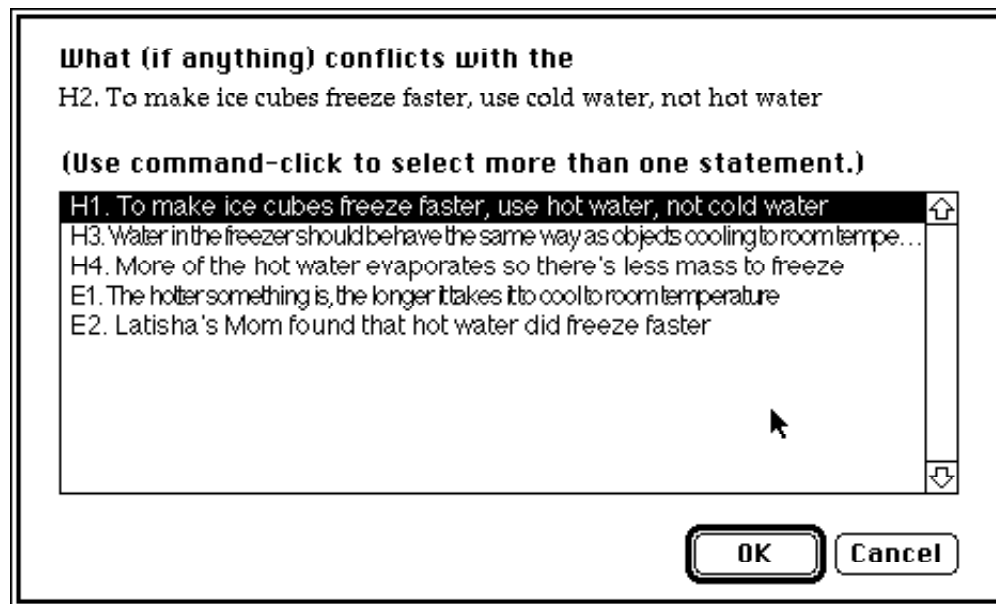


Figure 10. "Dialog box" for adding contradictions.

A "dialog box" will come up with a list of statements, and ask you to specify your contradictions (see Figure 10). Once again, you can select one or more statements by holding down the Command key when you point and click the mouse on a statement. Your contradictions will then show up in the contradictions window (see Figure 11).

To delete a contradictory statement, select the statement and then click the Delete Conflict button in the "**Contradictions** window."

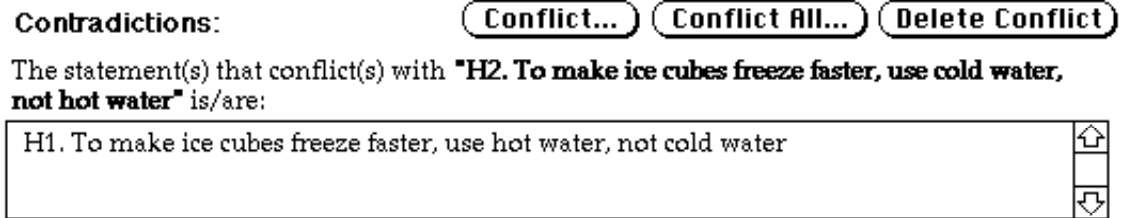


Figure 11. "Contradictions" window after the addition above.

Exercise 3

Add the following contradiction to your ice cubes argument by clicking on H2 and then on the Conflict... button, and selecting H1 from the dialog box.

H1 contradicts H2

OK, I've entered my argument. Now what?

Now you can run the simulation and see what the computer thinks. But first, you should rate how strongly you believe each of the statements you entered, so you have something to compare with the computer's evaluations.

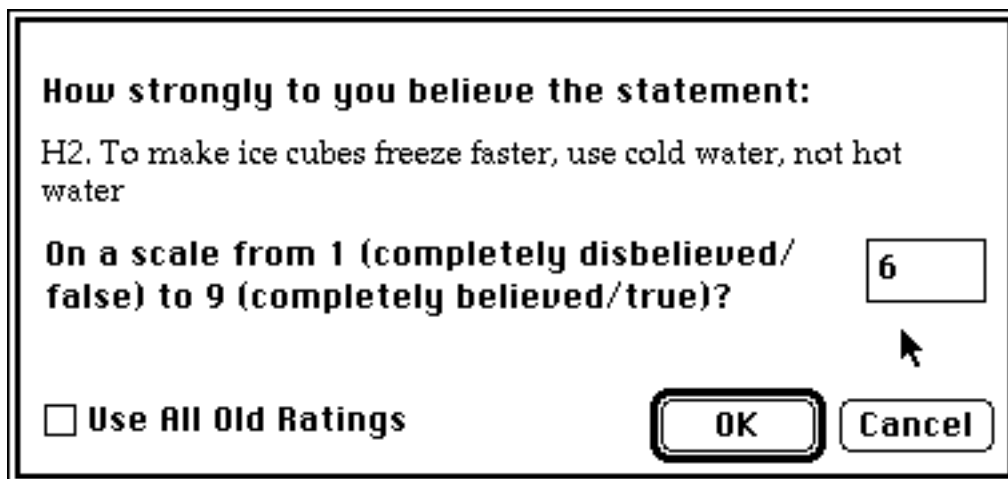
To do this, either select a statement that you want to rate and then click on the Rate... button, or just click the Rate All... button and let CONVINC ME ask you for ratings for all of the statements, one after the other (see Figure 12). Then enter your rating, on a scale from 1 (completely disbelieved), to 9 (completely believed), where 5 is "neutral," like so:

completely				neutral				completely
disbelieve/reject								believe/accept
1	2	3	4	5	6	7	8	9

If you're working with an argument that you saved earlier, and want to use ratings you offered previously—rather than re-rate all the statements, just check the **Use All Old Ratings** box in the ratings "dialog box." When you're done specifying your ratings, you can run the simulation to see what the computer thinks.

Exercise 4

Enter your believability ratings for the statements in the ice cubes argument.



How strongly to you believe the statement:

H2. To make ice cubes freeze faster, use cold water, not hot water

On a scale from 1 (completely disbelieved/false) to 9 (completely believed/true)?

Use All Old Ratings

6

OK Cancel

Figure 12. "Dialog box" for entering believability ratings.

Running the Simulation

You can change ECHO's numerical parameter settings before running the simulation, but it's not necessary. They're already set to some default "usual" values. We'll talk about more about these parameters later. To run the ECHO model, just go to the Simulation menu and select Run (see Figure 13). Later on, if you want to change the parameters, select Parameters... in the

Simulation menu. A "**Parameters** window" will appear at the lower right section of the screen (see Figure 14). If you change the parameters and then want to reset them to the original values, click on the **Use Default** button in the "**Parameters/Steps** window" (which would appear where the "steps" were).

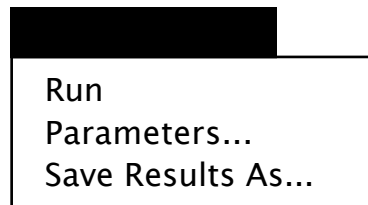


Figure 13. Simulation menu for running the model.

Parameters:

Explanation weight	<input type="text" value="0.03"/>
Conflict weight	<input type="text" value="0.06"/>
Evidence 'boost'	<input type="text" value="0.05"/>
Skepticism	<input type="text" value="0.04"/>

Figure 14. "Parameters window."

After the simulation has run, small "thermometer" icons will show up in the upper-right section of the screen (see Figure 15). Each statement that

you entered has one icon that represents it. The label for the statement, along with ECHO's evaluation (on the same 1 to 9 scale that you used to rate your belief in the statement), is below the icon. If you pass the mouse arrow over a thermometer, the text of the statement that it represents will show up in the "**Help/Message** window".

If the thermometer's "mercury" is above the half-way line (and gray), then ECHO generally "believes" (or accepts) your statement. The higher the mercury, the higher the activation, and the more ECHO accepts the statement. Similarly, if the mercury is below the half-way line (and black), ECHO "disbelieves" (or rejects) your statement to the degree that it's below the line. ECHO's activation ("temperature") for each statement is also displayed to the right of your **Ratings** after a simulation (see Figure 15).

Exercise 5

Run a simulation for your argument (choose Run from the Simulation menu).

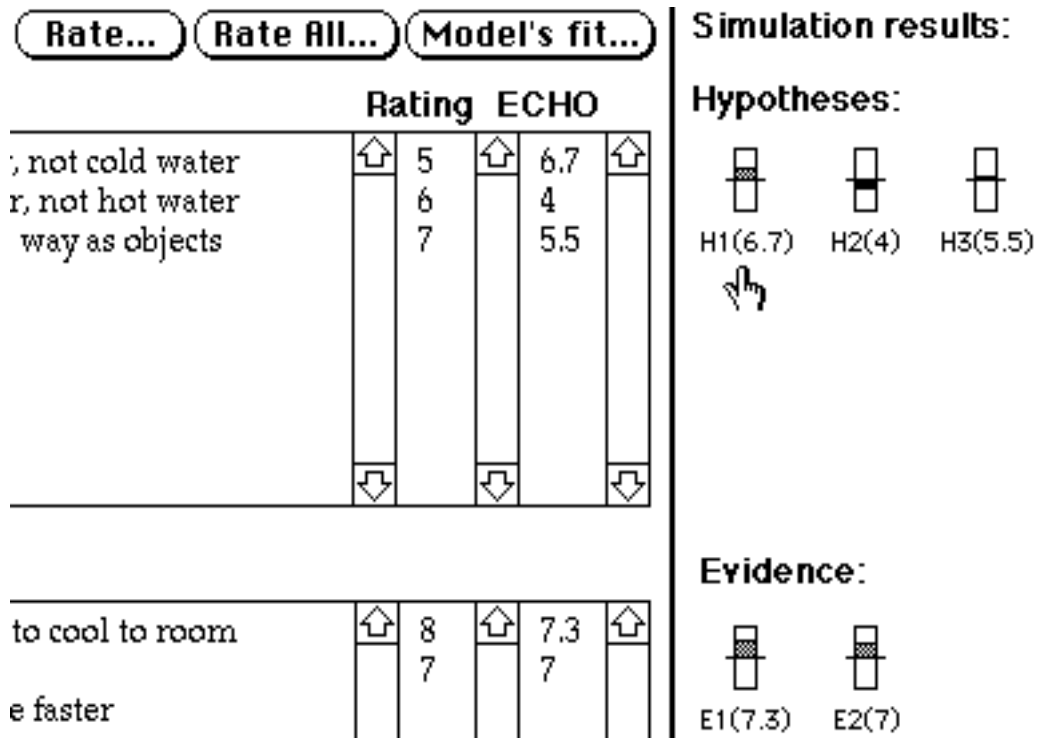


Figure 15. Parts of the "Statements" and "Activations" windows, showing "your" ratings and the model's final activations ("temperatures").

How do I compare my ratings to ECHO's activations?

Well, you can look at the two and see how they agree and disagree, and you can also request an overall measure of their agreement. You can do this by clicking on the **Models Fit...** button. (You have to have rated each statement for this to work, so if you haven't already, do so now.) ECHO will then compute an overall **correlation** between your ratings and ECHO's activations, and also tell you for which three statements your and ECHO's ratings disagree the most (see Figure 16). The higher the overall correlation, the more ECHO agrees with your ratings—based on your argument. (A negative correlation means that your ratings are actually *disagreeing* with ECHO's activations.) Table 1 shows the ranges of correlation values used to determine how related your ratings are to ECHO's activations overall.

Table 1. Determining the overall agreement between your evaluations and ECHO's.

<u>Correlation Range</u>		<u>Relation between your evaluations & ECHO's</u>	
-0.99	up to	-0.40	mostly opposite
-0.40	up to	-0.01	mildly opposed
		0.0	(unrelated)
0.01	up to	0.40	mildly related
0.40	up to	0.70	moderately related
0.70	up to	0.90	highly related
0.90	up to	0.99	almost identical

Exercise 6

How do your believability ratings compare to ECHO's? For what statements do your and ECHO's ratings differ the most? For which statements are they the most similar? Click on **Models Fit...** button. How well do your ratings agree with ECHO's overall? If you and ECHO didn't correlate as well as you thought you would, why might that be?

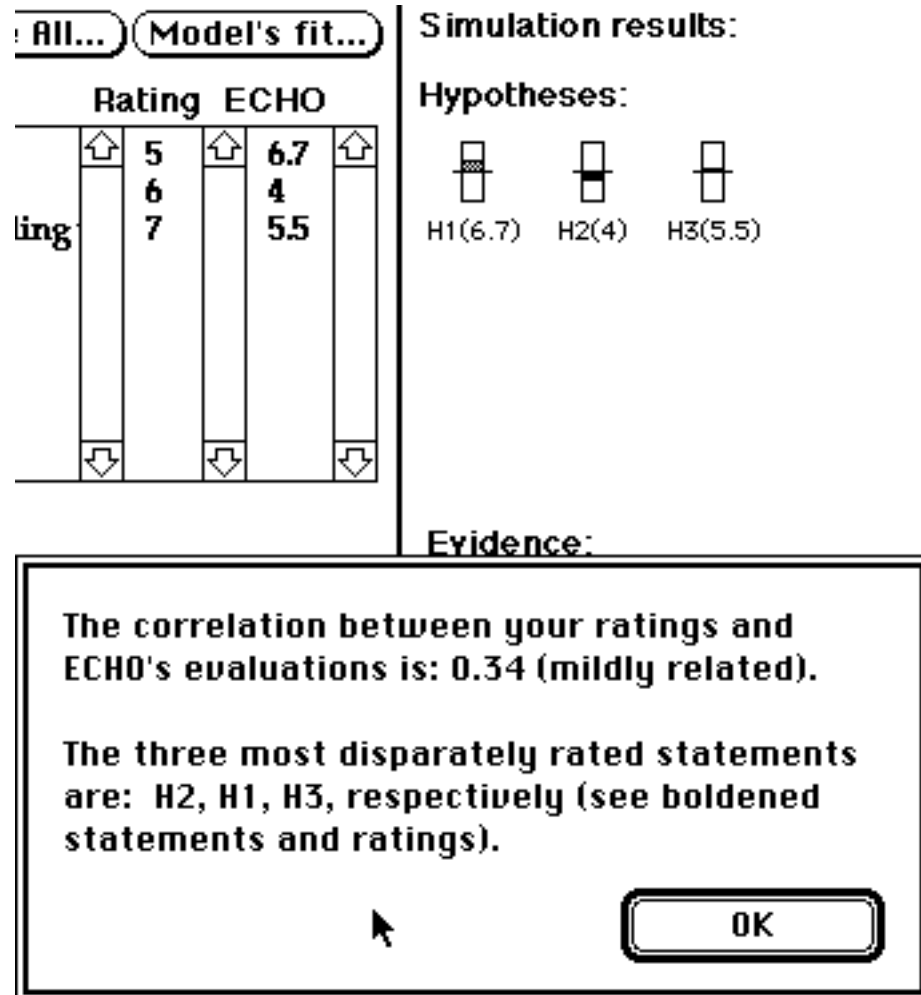


Figure 16. "Dialog box" to tell you how well you and the computer agree, and on which statements you disagree the most.

What if ECHO and I don't agree?

If you don't "convince" CONVINC ME the first time, that is, if ECHO doesn't agree with your evaluations, there are a few things you can try. For instance, look at the structure of your argument: Do you want to change it? Did you leave some explanations or contradictions out? Should some independent explanations be a joint explanation or vice versa? Look at your statements: Do you want to add or delete some? Do you want to change

some of your ratings? (Don't say that you believe something if you don't, just because ECHO "believes" it!)

Later on, you can even try change some of ECHO's numerical parameter settings to make ECHO better model your way of thinking. For example, if you think ECHO is being too "tolerant" compared to you, you might lower the **Explanation weight** and/or raise the **Contradiction weight**. If you think ECHO is not "tolerant" enough, you could raise the **Explanation weight** and/or lower the **Contradiction weight**. If you think ECHO isn't giving the proper weight to evidence, you could lower or raise the **Evidence 'boost'**. If you think ECHO is being too "skeptical", you could lower the **Skepticism weight**. If ECHO is not as skeptical as you, you might raise **Skepticism**.

It is possible that you may look at all of these things, make some changes, re-run the simulation, and ECHO still won't agree with you like you thought it would. That's okay, sometimes you just can't convince everyone, no matter how hard you try! But the important thing is that you think about your argument, reflect on it, and think about your own reasoning strategies.

Exercise 7

Implement at least one change to your ice cubes argument. Feel free to add, delete, or modify whatever statements, explanations, or contradictions that seem appropriate. (For example, you might add the information that since more of the hot water evaporates, there is less water to freeze, so it takes less time to freeze than the eventually-more-massive cold water.)

Exercise 8

Create an argument in CONVINC ME based on the "Rats" text from Unit 2. Run an ECHO simulation of your argument, and based on ECHO's feedback, make at least one change to your argument. Feel free to add, delete, or modify whatever statements, explanations, or contradictions that seem appropriate.

That's all for Unit 3! You may want to read through the summary glossary of terms on the next page. And see the following Appendix to learn more about how CONVINC ME evaluates your arguments.

Glossary

Argument: A system of beliefs that is generally more complex than one explanation/ contradiction, but less than that of a theory.

Belief: A hypothesis or piece of evidence.

Believability rating: Given a proposition, how strongly it is believed.

Confirmation bias: When one seeks to support certain arguments/beliefs in a biased fashion, without trying to disconfirm them.

Contradiction/Conflict: The relation between a pair of beliefs that are mutually exclusive or (at least) unlikely to both be true.

Disconfirmation: When one attempts to garner evidence that contradicts on (even favorite) theory.

Evidence: A belief that seems based on "objective-like" criteria; for example, an acknowledged common fact or statistic, or a reliable memory or observation.

Explanation: Something that shows how or why something happened. The coordination of beliefs such that some are accounted for (often causally) by others.

Hypothesis: One possible belief that explain/tells something of interest.

Joint Explanation: An explanation in which two or more beliefs together (vs. independently) explain a third belief.

Primacy bias: A tendency to give too much credence to early information.

Recency bias: A tendency to give too much credence to recent information.

Theory: A system of evidential and hypothetical beliefs that have a unifying theme.

Appendix: Some Principles That Underlie TEC and ECHO

(1) **Symmetry:** "Coherence and incoherence are symmetric relations." This means that if one belief explains (or conflicts with) another, the beliefs "send activation" back and forth to each other. (Cf., If I'm playing cards with you, then you're playing cards with me. If I'm not playing with you, then you're not playing with me.)

(2) **Explanation:** "A belief that explains a proposition coheres with it. Also, beliefs that jointly explain a proposition cohere with it, and cohere with each other. "Two or more beliefs that together explain a third belief are generally called "cohyphoteses" if they are both hypotheses (or sometimes "cobeliefs" if one or more is evidence). According to this principle, for example, cohyphoteses "send activation" to each other, as well as to the explained belief. (E.g., Todd singing and Mary singing jointly explains why it sounded like a duet, and send activation to "duet", as well as to each other.)

(3) **Simplicity:** "The plausibility of a proposition is inversely related to the number of explaining statements needed to explain it." The simpler the explanation, the more likely it will be believed. That is, lots of assumptions (or co-beliefs) are often counterproductive, compared to fewer assumptions.

(4) **Data Priority:** "Results of observations have an extra measure (boost) of acceptability." This means that acknowledged facts, memories, and observations carry more importance than "mere" hypotheses.

(5) **Contradiction:** "Contradictory hypotheses incohere." This means that beliefs that conflict with each other send "negative activation" (or "inhibition") to each other, like rival members of two different "gangs."

(6) **Competition:** "Competing beliefs (which explain the same evidence or hypotheses but are not themselves explanatorily related) incohere." This means that highly independent explainers of the same proposition conflict with each other, and hence send "negative activation" to each other, like rival gang members vying for the same turf. (E.g., If you hear a report that an evil dictator was shot, and later hear that he was stabbed, you might assume that

the two reports offer competing hypotheses.) This principle may be optionally invoked in a variant of ECHO, called ECHO2, which automatically infers inhibitory relationships between propositions that independently (i.e., not *jointly*, as in principle 2) explain a third proposition.

(7) **Acceptability:** "The acceptability of a proposition increases as it coheres more with other acceptable propositions, and *incoheres* more with *unacceptable* propositions." This basically says that how much a belief is believed is a function of who its friends and enemies are, and how much *they* are believed.

(8) **Overall Coherence:** "The overall coherence of a network of propositions depends on the local pairwise cohering of its propositions." This basically means that the goodness of a whole "neighborhood system" of beliefs is determined by the believability of its members and their relationships.

**APPENDIX E: Integrative Exercises
(Versions Used with *Convince Me* and
Written Groups)**

Exercises

For each exercise, subjects were given (a) a passage, and (b) a set of instructions. The passages are shown below, in the order that they were presented to the subjects. Instructions for *Convince Me* users differed slightly from those given to subjects who didn't use the system, and both sets of instructions are given below. Note that each subject was given a new copy of the instructions with each passage.

Passages

Exercise 1

Consider the following passage:

Wanda and Dave are walking through Pinetown one night, and both notice that an approaching teenager yawns when passing them.

Dave thinks that the teenager's yawn was an subconscious aggressive display. He learned in biology that humans are genetically close to apes, and ape studies suggest that apes engage in "threat yawns." In a group, dominant male apes yawn more—an action that shows off their long canine teeth—while subordinate apes more often cover their yawning mouths with their paws. He says that since Pinetown is a dangerous area, this would explain why the teenager yawned when passing them.

Wanda disagrees with Dave. She notes that people, as well as non-primates such as dogs, yawn when they are alone as well as in groups. She has read that yawning provides more oxygen to the brain and that the more oxygen, the more glucose we can burn for energy. She thinks that since it is late, the teenager is probably tired and yawned to get more oxygen to stay alert. She claims that the hypothesis that yawning is to increase oxygen also explains why it *seems* contagious — people in the same room are all just breathing the same stuffy air, and all need more oxygen.

Exercise 2

Consider the following passage:

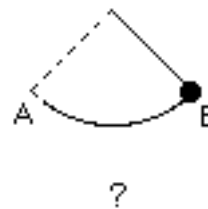
Two interns, Rafael and Sandra, are trying to decide which one of two patients has glumpis (as only one patient actually has it). Don has bloodshot eyes, they both have swollen eyelids, and Sylvia has a headache and a rash. Rafael thinks Don has been exposed to glumpis infection, so Don probably is the one with glumpis. Rafael also notes that if one assumes that Don has glumpis, that explains why Don has bloodshot eyes and swollen eyelids.

Sandra disagrees. Sandra thinks that Sylvia comes from a family that's rather susceptible to glumpis infections, so Sylvia probably has glumpis. Sandra also notes that if one assumes that Sylvia has glumpis, that explains why Sylvia has swollen eyelids, why Sylvia has a headache, and why Sylvia has a rash on her face.

Exercise 3

Consider the following situation:

A pendulum is swinging back and forth on a pendulum bob, and the bob is released exactly at the endpoint (E) of a swing (the farthest to the right it can go).



Draw as many plausible, alternative paths that you or someone else might think it could follow to the ground. Label your paths (e.g., P1, P2, P3, etc.).

Exercise 4

Consider the following passage:

Smith believes that abortion is wrong because fetuses are alive. Jones disagrees, saying the abortion is fine, because we as a society kill living things (e.g., for food) all the time.

Instructions

(One copy of the instructions was given with each exercise.)

Convince Me Group

- a) Using CONVINC ME, enter the hypotheses and evidence mentioned in the passage/situation.

- b) Using the Rate... or Rate All... buttons, rate how strongly you believe your statements.

- c) Using CONVINC ME, add any other plausible hypotheses or evidence that come to mind.

- d) Using CONVINC ME, enter the explanations and contradictions that seem appropriate.

- e) Using CONVINC ME, run an ECHO simulation of your argument.

For what statements do your and ECHO's ratings differ the most? Write them here:

For which statements are they the most similar? Write them here:

f) Click on the **Models Fit...** button. How well do your ratings agree with ECHO's overall? Write the correlation here:

g) Using **CONVINCE ME**, make any other revisions to your argument that seem appropriate.

h) Please revise your ratings in **CONVINCE ME**, if that seems appropriate.

Written Group

a) List the hypotheses mentioned in the passage/situation. Label them H1, H2, H3, etc.....:

b) List the evidence mentioned in the passage/situation. Label them E1, E2, E3, etc.....:

c) Below, please rate how strongly you believe the statements you listed in (a) and (b), on a scale from 1 (completely disbelieve/reject) to 9 (completely believe/accept). Write the label of the statement below, followed by the rating (e.g., H12: 3, E13: 5, etc...)

d) List (and label) any other plausible hypotheses that come to mind:

e) List (and label) any other evidence that comes to mind:

f) What statements *explain* what other statements? (You don't have to write the statements out. You can use the labels to say things like, "H5 explains E10")

g) What statements *contradict* what other statements? (You don't have to write the statements out. You can use the labels to say things like, "H1 contradicts H4")

h) Please revise your argument, if that seems appropriate, by writing any new hypotheses, evidence, explanations, or contradictions below. If you want to revise something that you wrote earlier, make your changes beside or below it in a **different color pen/pencil**.

i) Below, please revise your ratings, if that seems appropriate, for the statements you listed in (a), (b), (d), (e), and (h), on a scale from 1 (completely disbelieve/reject) to 9 (completely believe/accept). Write the label of the statement below, followed by the rating (e.g., H12: 3, E13: 5, etc...)

APPENDIX F: Post-test (Questions That Differ From the Pre-test Only)

Name _____ Date _____

Problem 1 (identical to pre-test)

Problem 2 (identical to pre-test)

[Note: statements b,d,f,h j l n, & p were not included in Study 1]

Problem 3 (isomorphic to pre-test)

(Tell your interviewer you're on problem 3. Then read on.) We have a rule in mind that governs a set of three numbers, and we'd like you to try to guess what it is. At any time, you can either:

- Propose sets of three numbers, to each of which your experimenter will reply "that set of numbers fits the rule" or "that set doesn't fit the rule," or
- State what you think the rule is, if you think you've figured it out.

Here is a set of three numbers to start: **5, 3, 1**

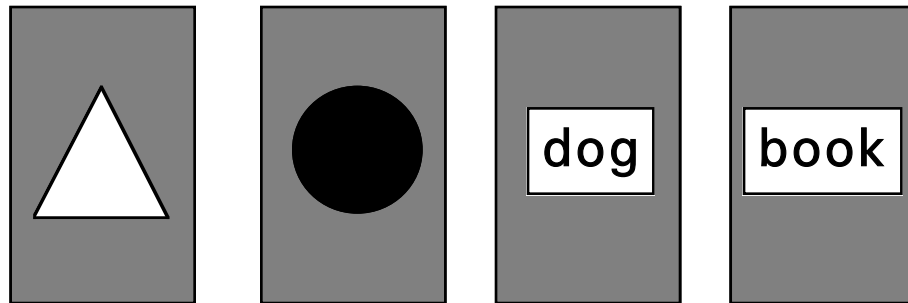
Try to use what you know about disconfirmation!

(Give this sheet to your interviewer. He or she will write down your proposed numbers and rules in the order you propose them.)

Problem 4 (identical to pre-test)**Problem 5 (isomorphic to pre-test)**

Using what you know about disconfirmation, consider the following situation:

You have four sheets of paper in front of you. Every sheet has a word on one side and a shape on the other side. Your task is to decide which sheets you need to turn over to determine the truth or falsity of this rule: *If there is a name of an animal on one side of a card, then there is a white triangle on the other side.*



What is the minimum number of cards you would need to turn over? What are they, if any?

Why did you choose the cards that you did?

Problem 6 (identical to pre-test, except following isomorphic passage used)

Consider the following passage:

Pat and Manfred try to decide where a native lives. In speaking, Uli used the open "oo", low back "ah", mid central "nya" and dropped "d" sounds. Pat thinks that Bawan is a big village, so Uli probably lives in Bawan. Pat also notes that if one assumes that Uli lives in Bawan, that explains why she used the open "oo" and why she used the low back "ah" sounds.

Manfred disagrees with Pat. Manfred believes that Uli lives in Woowee. Manfred says that he thinks that Woowee is a nearby village, so Uli probably lives in Woowee. Manfred also notes that if one assumes that Uli lives in Woowee, that explains why she used the low back "ah", why she used the mid central "nya", and why she used the dropped "d" sounds.

Problem 7 (identical to pre-test, except following isomorphic passage used)

Consider the following passage:

A child who has tested positive for the presence of HIV (AIDS virus) wishes to enter a pre-school. Are the other children in the school safe from becoming infected, or are they unsafe?

On one hand, casual transmission of the infection may not be possible. 95% of all childhood HIV cases are known to have contracted the infection from their mothers at or before birth, or from receiving blood transfusions. The Surgeon General has determined that transmission of the HIV infection by casual contact is extremely unlikely. And no mother of an HIV-positive child (who has contracted HIV through transfusion) has become infected from her child. Finally, the assumption that many viruses (such as HIV) can never be casually transmitted suggests that casual transmission of HIV is impossible. The unlikelihood of casual transmission of HIV would make it safe for the other children if the HIV-positive child were to attend the school.

On the other hand, HIV transmission through casual contact may indeed be possible. 5% of pediatric HIV cases are of unknown origin. In a number of hospitals, AIDS patients are separated from other patients. And the virus has been demonstrated to be present in saliva and tears. The possibility of casual transmission of HIV makes it unsafe for uninfected children to attend school with the HIV-positive child.

APPENDIX G: Exit Questionnaire

Name _____ Date _____

We want to get an idea of your impressions of the activities you've completed. (Don't worry, we'll pay you regardless of what you say here!). Ask the interviewer for a copy of units 1, 2, and 3 if he/she hasn't already given them to you.

1. What was your motivation for participating in this study (e.g., money, curiosity, interest in reasoning, boredom, ... other)?

2. How much did you learn from... What, if anything, did you learn?

<p>a. Unit 1 (Evidence, Hypotheses & Theories)?</p> <p>not much a lot</p> <p>1 2 3 4 5 6 7</p>	
<p>b. Unit 2 (Reasoning about Arguments)?</p> <p>not much a lot</p> <p>1 2 3 4 5 6 7</p>	
<p>c. Unit 3 (<i>Convince Me</i>)?</p> <p>not much a lot</p> <p>1 2 3 4 5 6 7</p>	
<p>d. using the <i>Convince Me</i> program?</p> <p>not much a lot</p> <p>1 2 3 4 5 6 7</p>	

<p>e. the exercises?</p> <p>not much a lot</p> <p>1 2 3 4 5 6 7</p>	
<p>f. the beginning and ending tests?</p> <p>not much a lot</p> <p>1 2 3 4 5 6 7</p>	

3. What did you like least about...

a. the readings?

b. the *Convince Me* program?

c. the exercises?

4. What did you like most about...

a. the readings?

b. the *Convince Me* program?

c. the exercises?

5. Do you have any suggestions for how to improve...

a. the readings?

b. the *Convince Me* program?

c. the exercises?

6. Any other comments or suggestions?

**APPENDIX H: New Unit 3, "Using
Convince Me," Extensively Revised for
the New Argument Diagram/Listing
Version of the Software**

Unit 3: Using *Convince Me*⁹

What is *Convince Me*?

Convince Me is a computer program to help you think about your own reasoning. The program lets you type in short, sentence-like statements: things you believe and are sure of, and beliefs/things you're not so sure of. Then you can tell the computer which ideas explain and contradict the other ideas (see Figure 1).

So what? Why do I need a computer for that?

You don't. But just as explaining something to another person can help you understand something, entering an argument into *Convince Me* can help you clarify your own beliefs. Also, just as people will often tell you what they agree and disagree with in your argument, *Convince Me* will, in a similar way, tell you which statements your argument helps to affirm or reject and which ones it leaves neutral, from the computer's point of view.

How does the computer know what to believe?

It doesn't, except for what you tell it. When you put a statement in the computer, you'll be asked whether it is a piece of evidence or a hypothesis. Decide carefully, since the computer gives more weight to all pieces of evidence and then tries to figure out which hypotheses and evidence "hang together" best. **The computer doesn't understand the meanings of the statements that you type in.** It just tries to figure out which statements to believe on the basis of your argument—by what you tell it about what

⁹*Convince Me* was developed by the ECHO Educational Program (EEP), at the University of California, Berkeley. © 1993 and 1994 University of California.

contradicts what, and what explains what. *Convince Me* uses a computer program called ECHO to do this.

What is "ECHO"?

ECHO is a computer model based on a theory called the "Theory of Explanatory Coherence" (TEC). The next section describes TEC and ECHO in more detail.

CM (big).u3.0 cubes

—Ratings— Add... Edit... Delete... Rate... Rate All... Model's fit...

You	ECHO	Hypotheses:
5	5.4	H1. To make ice cubes freeze faster, use hot water, not cold water
6	5.3	H2. To make ice cubes freeze faster, use cold water, not hot water
7	5.6	H3. Water in the freezer should behave the same way as objects cooling to room

You	ECHO	Evidence:
8	7	E1. The hotter something is, the longer it takes it to cool to room temperature
7	6.3	E2. Latisha's Mom found that hot water did freeze faster

Explanations: Explain... Delete Explanation

H3. Water in the freezer should behave the same way as objects cooling to room temperature *AND*
E1. The hotter something is, the longer it takes it to cool to room temperature

Explain(s) why: "H2. To make ice cubes freeze faster, use cold water, not hot water"

Contradictions: Conflict... Delete Conflict

H1. To make ice cubes freeze faster, use hot water, not cold water

Conflict(s) with: "H2. To make ice cubes freeze faster, use cold water, not hot water"

Help: E1. The hotter something is, the longer it takes it to cool to room temperature (click & drag node to rearrange graph) **File:**

Graph and simulation results: Hide links

All Explanations & Contradictions:

H1 explains E2
explains
H3 E1 jointly explain H2

H1 contradicts H2

Steps for using CONVINC ME:

1. Enter hypotheses and evidence
2. Enter explanations and contradictions.
3. Rate the believability of your statements.
4. Run the ECHO simulation.
5. Compare your evaluations to ECHO's.
6. (optional) Make changes based on ECHO's feedback.

The correlation between your ratings and ECHO's evaluations is: 0.34 (mildly related).

The three most disparately rated statements are: H2, H1, H3, respectively (see boldened statements).

Your statement:

More of the hot water evaporates so there's less mass to freeze

Check all that apply:

- Acknowledged fact or statistic
- Observation or memory
- One possible inference, opinion, or view
- Some reasonable people might disagree

Select one:

- Evidence E3 Reliability, if evidence? (from 1, poor, to 3, good)
- Hypothesis H4

OK

Cancel

Figure 1. Adding a belief about the speeds at which water of different initial temperatures freezes (bottom) in response to *Convince Me's* feedback (middle).

TEC and ECHO (you can skim the next two pages if you like)

The Theory of Explanatory Coherence (TEC) attempts to account for how people decide the plausibility of beliefs asserted in an explanation or argument. The theory is based on a few "hall of fame" **principles of reasoning**, such as:

- 1) The believability of an idea generally increases with increasing simplicity. In other words, making lots of (that is, joint) assumptions is often counterproductive, compared to making fewer assumptions.
- 2) People tend to believe statements when there is more evidence to support them.
- 3) We are more likely to believe something that doesn't conflict or compete with other things we strongly believe.

Etc. To learn more about TEC's principles, see the Appendix.

ECHO is a computer model based on TEC. In ECHO, arguments are represented as **networks of nodes** (like knots in a net). A hypothesis or piece of evidence is represented by a node, and explanatory or contradictory relations are represented by **links** between nodes. Hypothesis evaluation is treated as the satisfaction of **constraints** determined from the explanatory relations (that is, explanations and/or contradictions), TEC's principles, and from a few **numerical parameters**. Given a network of statements and relations between them, node activations are updated in parallel using a simple "**connectionist**" settling scheme. When the network of statements settles (or stabilizes), the nodes representing the most mutually coherent hypotheses and evidence are active, and the nodes representing inconsistent rivals are deactivated.

For example, suppose Chris says:

"Some people think that all animals (including humans) were created in their present form, about 5000 years ago. Others believe that animals evolved from earlier life slowly, over

millions of years. Both beliefs explain why animals exist. However, only the latter, evolutionary, hypothesis explains why transitions between forms in the fossil records appear to be gradual, and why scientists have found some fossils they estimate are over a million years old.

This could be represented in ECHO as:

hypothesis H1: "Animals were created in their present form about 5000 years ago."

hypothesis H2: "Animals evolved from earlier life over millions of years."

evidence E1: "Animals exist."

evidence E2: "Transitions between forms in the fossil records are gradual."

evidence E3: "Scientists have dated some fossils at over a million years old."

H1 competes with H2.

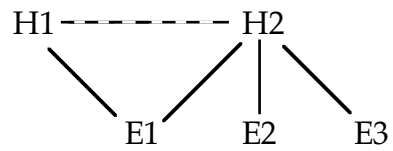
H1 explains E1.

H2 explains E1.

H2 explains E2.

H2 explains E3.

Or, in graphical network form (where solid lines represent explanatory links, and the dashed line represents a competing/contradictory link):



Given a scenario such as this, ECHO generates a numerical value for each statement that indicates how much it believes the statement. In general, the more positive the value, the more ECHO "believes" the statement; the more negative the value, the more ECHO "disbelieves" the statement. In this case, ECHO believes H2 over H1 since H2 explains more of the evidence.

**Please STOP skimming (and start reading thoroughly again)
from here on!**

Getting Help from *Convince Me*

If you have any questions about *Convince Me* or ECHO, select **About Convince Me...** or **About ECHO...** in the **Help** menu. To see a glossary of terms, select **Glossary...** in the **Help** menu. A glossary is also included at the end of this document. To see a summary list of steps about how to use *Convince Me*, select the **Steps...** item. When the steps are displayed, a checkmark (✓) will show up beside **Steps...** in the menu. This document will go through these steps in detail.

You can also use the **Help** menu to turn **Help Mode** on or off (the default is on). When help mode is on, a checkmark (✓) will show up beside **Help Mode** in the menu, and, when you pass the mouse cursor over parts of the software, messages will show up in the "**Help** window" located at the bottom of the screen (see Figure 2).

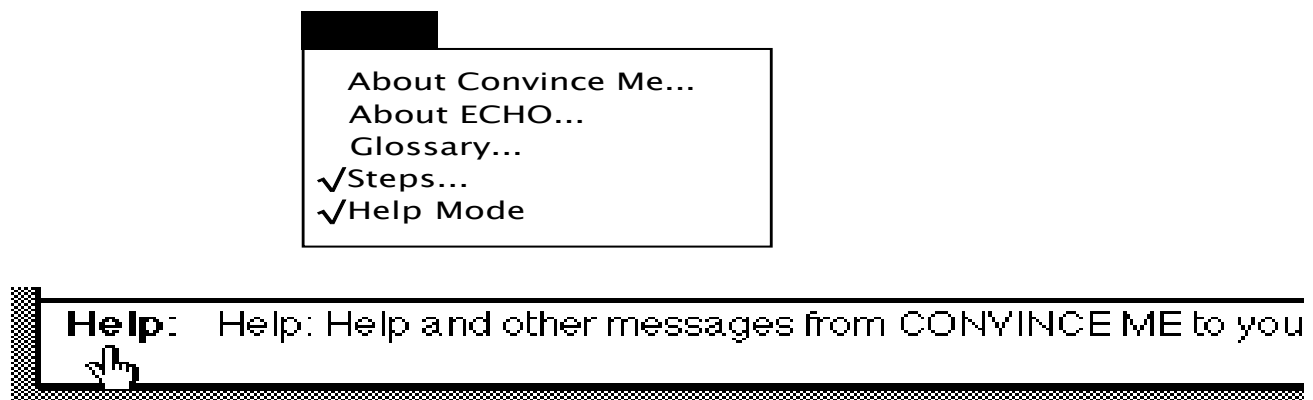


Figure 2. The Help menu and "Help window".

Entering an argument

The Argument menu lets you create a new argument, load an existing argument, or save your argument (see Figure 3).

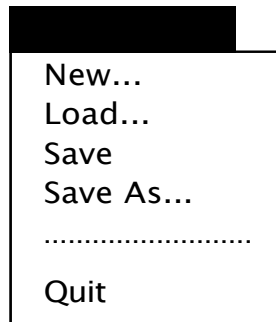


Figure 3. The Argument menu

—Ratings—
You ECHO

Add... Edit... Delete Rate... Rate All... Model's fit...

Hypotheses:

		H1. To make ice cubes freeze faster , use hot water, not cold water	↑
		H2. To make ice cubes freeze faster, use cold water , not hot water	
		H3. Water in the freezer should behave the same way as objects cooling to room	↓

You ECHO **Evidence:**

		E1. The hotter something is, the longer it takes it to cool to room temperature	↑
		E2. Latisha's Mom found that hot water did freeze faster	
			↓

Figure 4. "Statements window", with part of the "Ice Cubes" argument we saw earlier.

When you want to enter statements for your argument, click the Add... button in the upper left section of the *Convince Me* screen (see Figure 4). A "dialog box" will then ask you what statement you would like to add (see Figure 5). It will also ask you to check one or more of the boxes to help determine if the statement is a hypothesis or a piece of evidence, and it also asks you to explicitly decide which one it is. (You may check no boxes if none really apply at all.) If the statement is a piece of evidence, *Convince Me* also wants to know how "reliable" you think it is, on a scale from 1 (not very reliable) to 3 (very reliable):

Reliability (if evidence) is: poor fair good
 1 2 3

Your statement:
Latisha's Mom found that hot water did freeze faster

Check all that apply:
 Acknowledged fact or statistic
 Observation or memory
 One possible inference, opinion, or view
 Some reasonable people might disagree

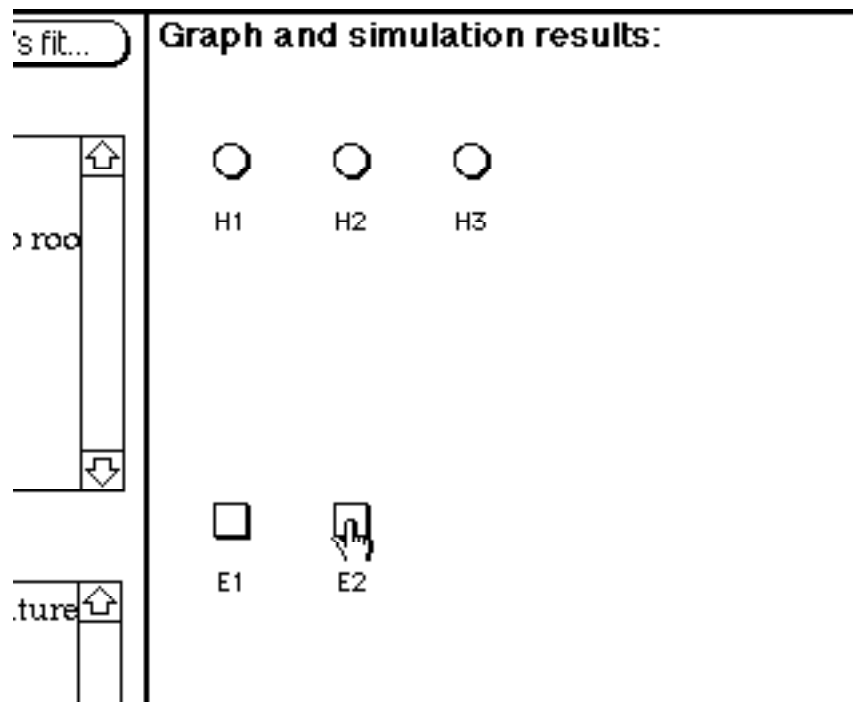
Select one:
 Evidence E2 Reliability, if evidence? (from 1, poor, to 3, good) 1.5
 Hypothesis H4

OK Cancel

Figure 5. "Dialog box" to add or edit a statement.

Each statement that you add is represented by an icon in the "**Graph and simulation results** window" located in the upper right of the screen. As you enter your propositions, a round icon appears for every hypothesis, and

a square icon appears for every piece of evidence (see Figure 6). The label for the statement is below the icon, and if you pass the mouse arrow over an icon, the text of the statement that it represents will show up in the "Help window".



Help: E2. Latisha's Mom found that hot water did freeze faster (click & c

Figure 6. Proposition icons in the "Graph and simulations results window" and text of proposition E2 in the "Help window."

If you want to change the text of a statement, or reclassify it as hypothesis or evidence or vice versa, click on the statement you want to modify and then click the Edit... button. If you want to delete a statement, select the statement and then click the Delete button (see Figure 4). After you've entered some statements, you can specify some explanations and contradictions among them.

Exercise 1

Create a new argument by selecting **New** from the **Argument** menu. Using the **Add...** button, add the following hypotheses and evidence to your argument (from the "Ice Cubes" argument in Unit 2). Don't specify any explanations and contradictions yet.

Hypotheses:

- To make ice cubes freeze faster, use hot water, not cold water (H1).
- To make ice cubes freeze faster, use cold water, not hot water (H2).
- Water in the freezer should behave the same way as objects cooling to room temperature (H3).

Evidence:

- The hotter something is, the longer it takes it to cool to room temperature (E1).
- Latisha's Mom found that hot water did freeze faster (E2).

Adding and deleting explanations

To create an explanation, first select a statement in the "**Statements** window" that you want to explain (e.g., if you want to explain the statement, "To make ice cubes freeze faster, use cold water, not hot water," then click on it). Then click on the **Explain...** button in the "**Explanations** window" (see Figure 7).

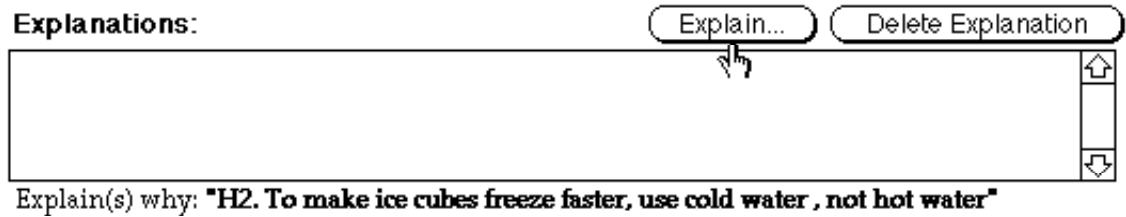


Figure 7. "Explanations window" before adding explanations.

A "dialog box" will then come up with a list of statements, and ask you to specify your explanations (see Figure 8). You can select multiple statements by holding down the Command key when you click on a statement. The Command key is the one that has the apple (🍏) on it, between the Option key and the spacebar.

To delete an explanation, select it and then click the Delete Explanation button in the "Explanations window."

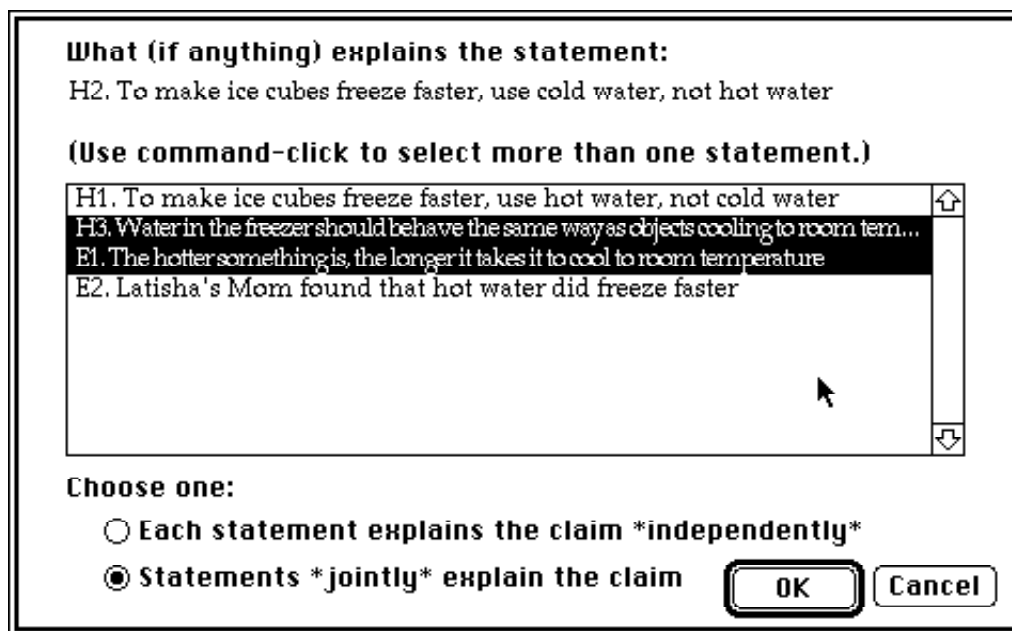


Figure 8. "Dialog box" for adding explanations.

When you enter an explanation, the computer will ask you if the explanations that you select *independently* or *jointly* explain your claim.

If you click "**Each statement explains the claim *independently***", this means that each statement explains your claim *on its own*, i.e., "<statement one> explains the claim", and "<statement two> explains the claim..." etc. (E.g., That Todd was singing explains why music was coming from the room, and that Mary was singing also—independently—explains why music was coming from the room.)

Click "**Statements *jointly* explain the claim**" if the statements *together, in conjunction*, explain the claim" (that is, <statement one> alone doesn't explain the claim, but together with the other statement(s) you get a proper explanation; e.g., Todd singing and Mary singing jointly explains why it sounded like a duet).

Your explanations then appear in the explanations window (see Figure 9).

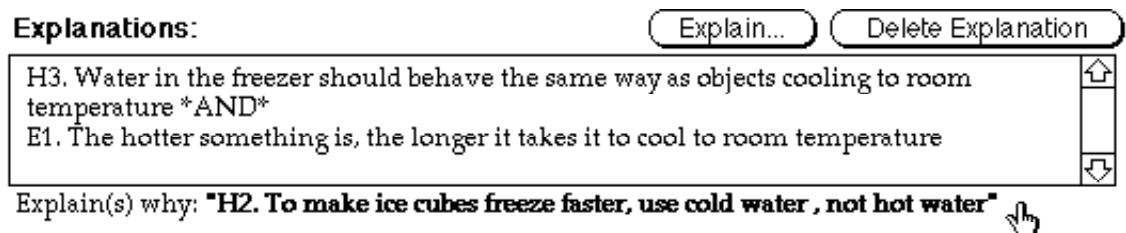


Figure 9. "Explanations window" after the first addition from Exercise 2.

Exercise 2

Add the following explanations to your ice cubes argument. To add the first explanation, click on H2 and then on the Explain... button, select E1 and H3 from the dialog box, and click on Statements **jointly** explain the claim. To add the second explanation, click on E2 and then on the Explain... button, select H1 from the dialog box, and click on Each statement explains the claim **independently**.

E1 and H3 jointly explain H2

H1 explains E2

Adding and deleting contradictions

To specify a contradiction, first select a statement in the "**Statements** window" that you want to contradict (e.g., if you want to specify what conflicts with the statement "To make ice cubes freeze faster, use cold water, not hot water," then click on it). Then click on the Conflict... button in the "**Contradictions** window" (see Figure 10).

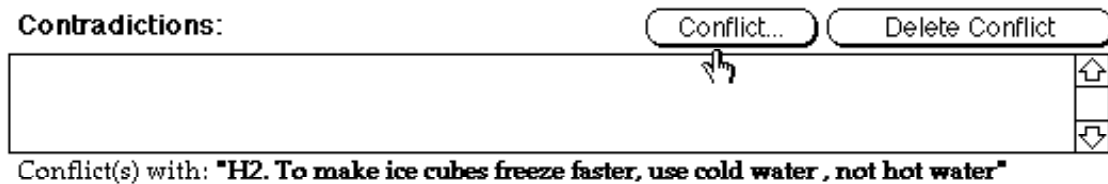


Figure 10. "Contradictions window" before adding contradictions.

A "dialog box" will come up with a list of statements, and ask you to specify your contradictions (see Figure 11). Once again, you can select one or more statements by holding down the Command (⌘) key when you point and click the mouse on a statement. Your contradictions will then show up in the contradictions window (see Figure 12).

To delete a contradiction, select the statement and then click the Delete Conflict button in the "**Contradictions** window."

Exercise 3

Add the following contradiction to your ice cubes argument by clicking on H2 and then on the Conflict... button, and selecting H1 from the dialog box.

H1 contradicts H2

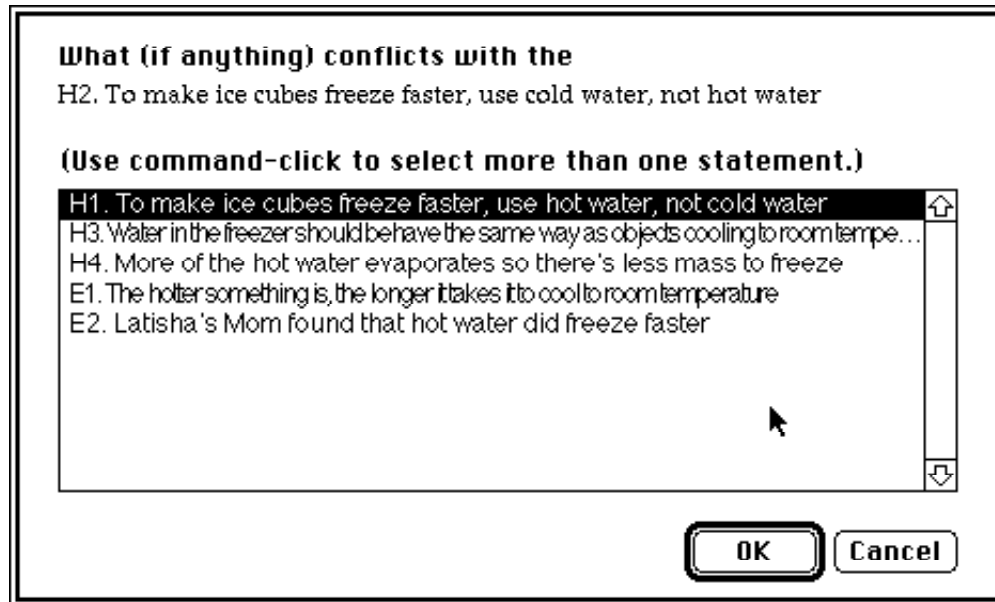


Figure 11. "Dialog box" for adding contradictions.

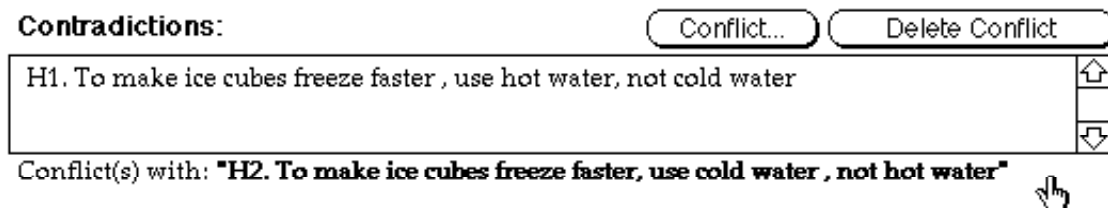


Figure 12. "Contradictions window" after the addition from Exercise 3.

Reviewing explanations and contradictions

If you wish to quickly review the explanations and contradictions in your argument, you can use the “**Listing** window” which is located at the bottom right-center of the screen (see Figure 13). You can scroll through a complete list of your explanations and contradictions. They appear in the order in which you entered them.

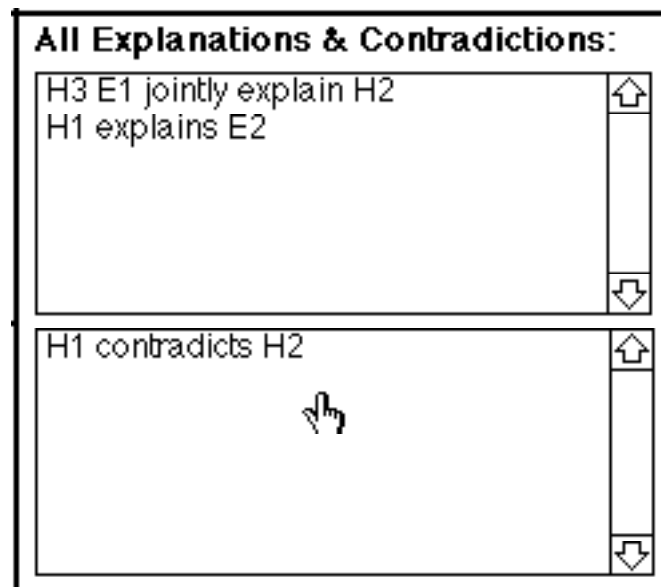


Figure 13. “Listing window” with all explanations and contradictions.

Exercise 4

How are joint explanations represented in the “**All Explanations & Contradictions Listing** window” for the “Ice Cubes” argument? Which of the explanations entered so far are joint explanations (write them below)?

Building a diagram of your argument

Convince Me will draw a diagram of your argument in the “**Graph and simulation results** window” which is located at the upper right of the screen (see Figure 14). You can even rearrange the diagram so that it makes sense to you (see Figure 15).

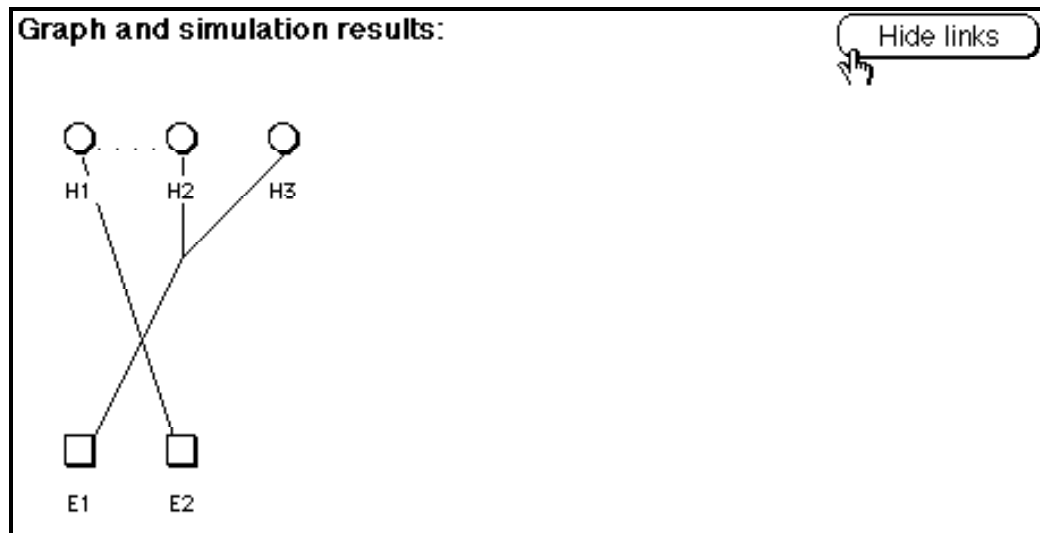


Figure 14. “Graph window” for Ice Cubes argument with original arrangement of icons.

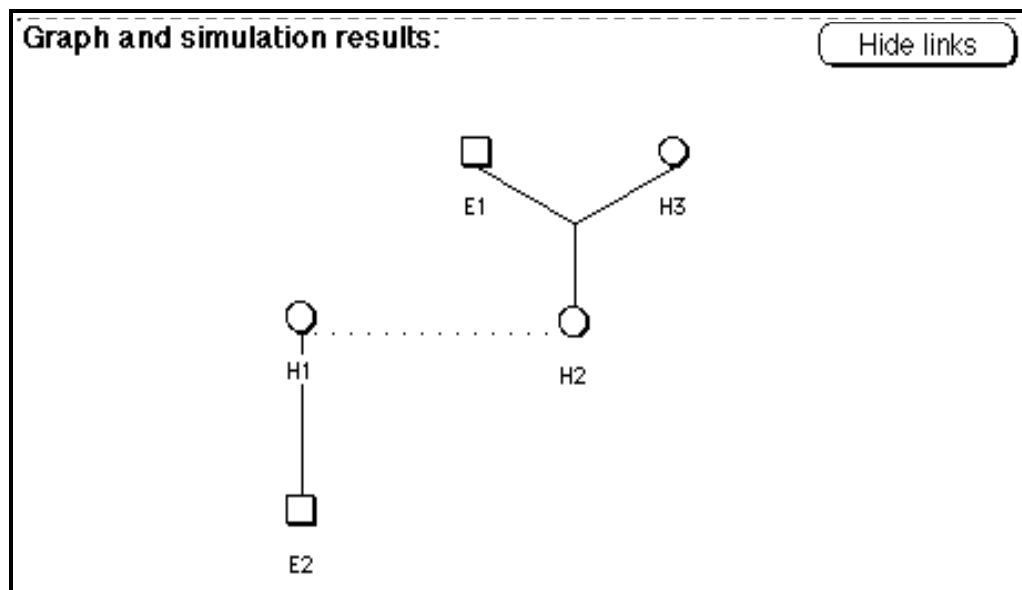


Figure 15. “**Graph** window” for Ice Cubes argument with icons rearranged to form a more meaningful diagrammatic representation of the argument.

You can move the icons by clicking on them with the mouse pointer and slowly dragging them across the screen while holding down the mouse button. When you click on the button labeled **Show Links**, the explanatory links you have entered will be drawn with solid lines and the contradictory links will be drawn with dotted lines.

When you click the **Show Links** button, its label changes to **Hide Links**. You may find it convenient to hide the links again whenever you want to rearrange the icons in the diagram so that *Convince Me* doesn't take extra time redrawing the argument every time. If the links are showing, the diagram will be updated automatically whenever you add a new explanation or contradiction.

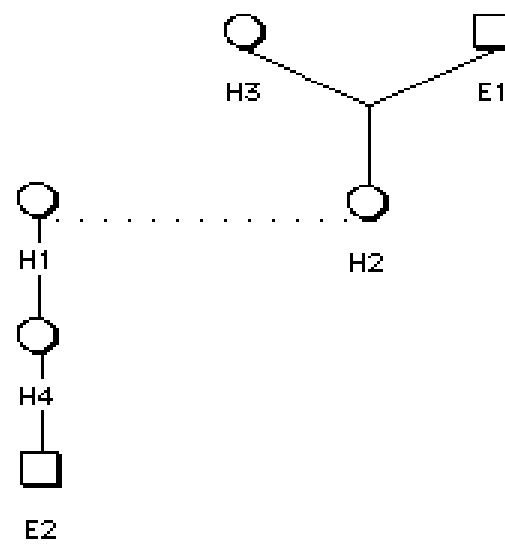
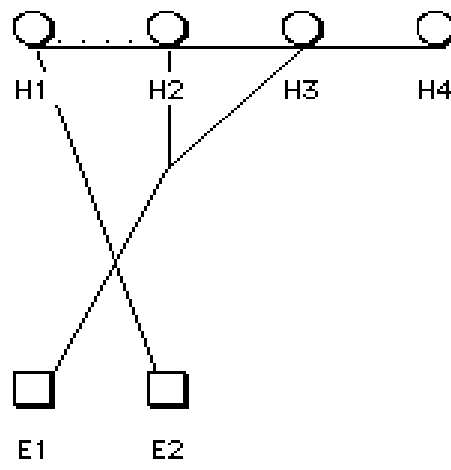
Sometimes an explanation or contradiction link (solid or dotted line), or even a proposition icon, on the diagram may be hidden due to the positioning of the proposition icons. If you add an explanation, contradiction, or statement and you notice that it does not appear on the diagram, rearrange the icons so that all the links and icons are displayed (see example in Figure 16).

Exercise 5

Arrange the icons for the "Ice Cubes" argument as pictured in Figure 15 and click the **Show Links** button.

Exercise 6

The diagram that you just built is arranged such that the propositions that explain something else are positioned *above* the propositions they are explaining. Also, the propositions supporting a single hypothesis are positioned nearby, whereas propositions supporting more than one hypothesis could be positioned centrally. Does this seem to be a good way to view the argument? What other features of the diagram may help you to "see" the argument better? Feel free to rearrange the diagram so that it represents the argument in a form meaningful to you.



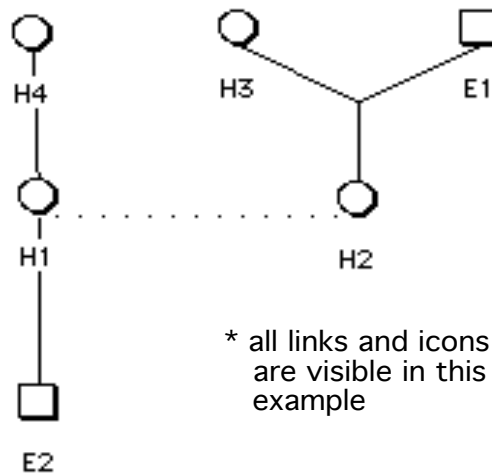
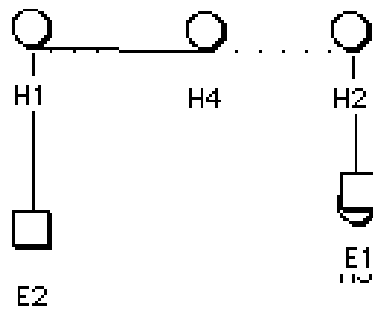


Figure 16. Four different graphs with the same explanations, and contradictions. Notice how the arrangement of icons can "hide" certain links between propositions and even hide other icons.

OK, I've entered my argument. Now what?

Now you can run the simulation and see what the computer thinks. But first, you should rate how strongly you believe each of the statements you entered, so you have something to compare with the computer's evaluations.

To do this, either select a statement that you want to rate and then click on the **Rate...** button, or just click the **Rate All...** button and let *Convince Me* ask you for ratings for all the statements, one after the other (see Figure 17). Then enter your rating, on a scale from 1 (completely disbelieved), to 9 (completely believed), where 5 is "neutral," like so:

completely disbelieve/reject	neutral	completely believe/accept
1 2 3	4 5 6	7 8 9

If you're working with an argument that you saved earlier, and want to use ratings you offered previously—rather than re-rate all the statements, just check the **Use All Old Ratings** box in the ratings "dialog box." When you're done specifying your ratings, you can run the simulation to see what the computer thinks.

How strongly to you believe the statement:

H2. To make ice cubes freeze faster, use cold water, not hot water

On a scale from 1 (completely disbelieved/false) to 9 (completely believed/true)?

Use All Old Ratings

Figure 17. "Dialog box" for entering believability ratings.

Exercise 7

Enter your believability ratings for the statements in the "Ice Cubes" argument.

Running the Simulation

You can change ECHO's numerical parameter settings before running the simulation, but it's not necessary. They're already set to some default "usual" values. We'll talk more about these parameters later. To run the ECHO model, just go to the Simulation menu and select Run (see Figure 18). Later on, if you want to change the parameters, select Parameters... in the Simulation menu. A "Parameters window" will appear at the lower right section of the screen (see Figure 19). If you change the parameters and then want to reset them to the original values, click on the Use Default button in the "Parameters window" (which would appear where the "steps" were).

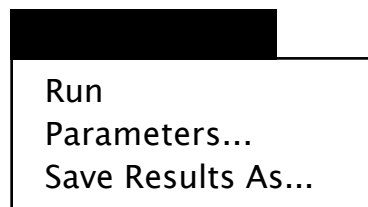


Figure 18. Simulation menu for running the model.

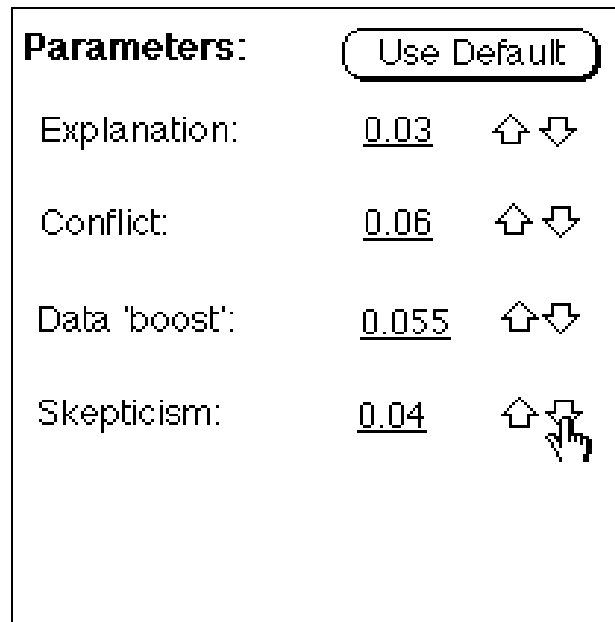


Figure 19. "Parameters window."

After the simulation has run, the icons in the **"Graph and simulation results window"** will register "thermometer" readings (see Figure 20). If the thermometer's "mercury" is above the half-way line, then ECHO generally "believes" (or accepts) your statement. The higher the mercury, the higher the activation, and the more ECHO accepts the statement. Similarly, if the mercury is below the half-way line, ECHO "disbelieves" (or rejects) your statement to the degree that it's below the line. ECHO's activation ("temperature") for each statement is also displayed to the right of your Ratings after a simulation (see Figure 20).

Exercise 8


Run a simulation for your argument (choose Run from the Simulation menu).

Add...

Edit

—Ratings—

You ECHO **Hypotheses:**

You	ECHO	
5	6.7	H1. To make ice c
6	4	H2. To make ice c
7	5.5	H3. Water in the
		

You ECHO **Evidence:**

You	ECHO	
8	7.3	E1. The hotter so
7	7	E2. Latisha's Mor

Graph and simulation resu

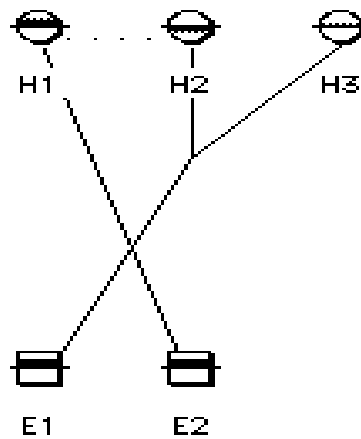


Figure 20. Parts of the "Statements" and "Graph/Activations" windows, showing your ratings and the model's final activations ("temperatures").

How do I compare my ratings to ECHO's activations?

Well, you can look at the two and see how they agree and disagree, and you can also request an overall measure of their agreement. You can do this by clicking on the **Models Fit...** button. (You have to have rated each statement for this to work, so if you haven't already, do so now.) ECHO will then compute an overall **correlation** between your ratings and ECHO's activations, and also tell you for which three statements your and ECHO's ratings disagree the most (see Figure 21). The higher the overall correlation, the more ECHO agrees with your ratings—based on your argument. (A negative correlation means that your ratings are actually *disagreeing* with

ECHO's activations.) Table 1 shows the ranges of correlation values used to determine how related your ratings are to ECHO's activations overall.

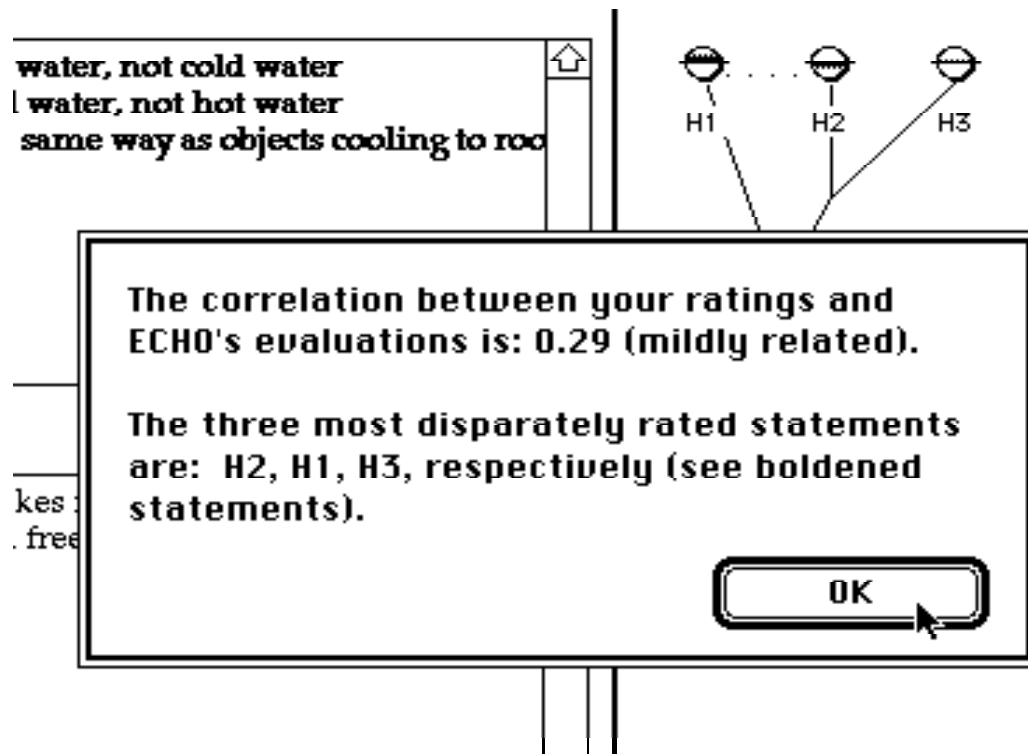


Figure 21. "Dialog box" to tell you how well you and the computer agree, and on which statements you disagree the most.

Table 1. Determining the overall agreement between your evaluations and ECHO's.

<u>Correlation Range</u>		<u>Relation between your evaluations & ECHO's</u>	
-0.99	up to	-0.40	mostly opposite
-0.40	up to	-0.01	mildly opposed
		0.0	(unrelated)
0.01	up to	0.40	mildly related
0.40	up to	0.70	moderately related
0.70	up to	0.90	highly related
0.90	up to	0.99	almost identical

Exercise 9

How do your believability ratings compare to ECHO's? For what statements do your and ECHO's ratings differ the most? For which statements are they the most similar? Click on Models Fit... button. How well do your ratings agree with ECHO's overall? If you and ECHO didn't correlate as well as you thought you would, why might that be?

What if ECHO and I don't agree?

If you don't "convince" *Convince Me* the first time, that is, if ECHO doesn't agree with your evaluations, there are a few things you can try. For instance, look at the structure of your argument: Do you want to change it? Did you leave some explanations or contradictions out? Should some independent explanations be a joint explanation or vice versa? Look at your

statements: Do you want to add or delete some? Do you want to change some of your ratings? (Don't say that you believe something if you don't, just because ECHO "believes" it!)

Later on, you can even try changing some of ECHO's numerical parameter settings to make ECHO better model your way of thinking. For example, if you think ECHO is being too "tolerant" compared to you, you might lower the **Explanation weight** and/or raise the **Contradiction weight**. If you think ECHO is not "tolerant" enough, you could raise the **Explanation weight** and/or lower the **Contradiction weight**. If you think ECHO isn't giving the proper weight to evidence, you could lower or raise the **Evidence 'boost'**. If you think ECHO is being too "skeptical," you could lower the **Skepticism weight**. If ECHO is not as skeptical as you, you might raise **Skepticism**.

It is possible that you may look at all of these things, make some changes, re-run the simulation, and ECHO still won't agree with you like you thought it would. That's okay, sometimes you just can't convince everyone, no matter how hard you try! But the important thing is that you think about your argument, reflect on it, and think about your own reasoning strategies.

Exercise 10

Implement at least one change to your ice cubes argument. Feel free to add, delete, or modify whatever statements, explanations, or contradictions that seem appropriate. (For example, you might add the information that since more of the hot water evaporates, there is less water to freeze, so it takes less time to freeze than the eventually-more-massive cold water.) Rearrange the diagram to reflect your changes to the argument.

Exercise 11

Create an argument in *Convince Me* based on the "Rats" text from Unit 2 (the argument is reproduced in Table 2 below). Run an ECHO simulation of your argument, and based on ECHO's feedback, make at least one change to your argument. Feel free to add, delete, or modify whatever statements, explanations, or contradictions that seem appropriate.

Table 2. Text of "Rats" exercise from Unit 2.

A UC Berkeley researcher believed that interesting, educational experiences in early life lead to larger brains. She found that rats raised alone in the empty cages had smaller brains than the rats raised together in the interesting environment. Based on this experiment, she concluded that children who have interesting, educational experiences in preschools will grow up to be more intelligent adults than children who do not attend preschool.

A preschool teacher disagreed with the researcher. She said that the rat experiment could not be used to explain the advantages of attending preschool.

That's all for Unit 3! You may want to read through the summary glossary of terms on the next page. And see the following Appendix to learn more about how *Convince Me* evaluates your arguments.

Glossary

Argument: A system of beliefs that is generally more complex than one explanation/ contradiction, but less than that of a theory.

Belief: A hypothesis or piece of evidence.

Believability rating: Given a proposition, how strongly it is believed.

Confirmation bias: When one seeks to support certain arguments/beliefs in a biased fashion, without trying to disconfirm them.

Contradiction/Conflict: The relation between a pair of beliefs that are mutually exclusive or (at least) unlikely to both be true.

Disconfirmation: When one attempts to garner evidence that contradicts a (even favorite) theory.

Evidence: A belief that seems based on "objective-like" criteria; for example, an acknowledged common fact or statistic, or a reliable memory or observation.

Explanation: Something that shows how or why something happened. The coordination of beliefs such that some are accounted for (often causally) by others.

Hypothesis: One possible belief that explain/tells something of interest.

Joint Explanation: An explanation in which two or more beliefs together (vs. independently) explain a third belief.

Primacy bias: A tendency to give too much credence to early information.

Recency bias: A tendency to give too much credence to recent information.

Theory: A system of evidential and hypothetical beliefs that have a unifying theme.

Appendix: Some Principles That Underlie TEC and ECHO

(1) **Symmetry:** "Coherence and incoherence are symmetric relations." This means that if one belief explains (or conflicts with) another, the beliefs "send activation" back and forth to each other. (Cf., If I'm playing cards with you, then you're playing cards with me. If I'm not playing with you, then you're not playing with me.)

(2) **Explanation:** "A belief that explains a proposition coheres with it. Also, beliefs that jointly explain a proposition cohere with it, and cohere with each other. "Two or more beliefs that together explain a third belief are generally called "cohyphoteses" if they are both hypotheses (or sometimes "cobeliefs" if one or more is evidence). According to this principle, for example, cohyphoteses "send activation" to each other, as well as to the explained belief. (E.g., Todd singing and Mary singing jointly explains why it sounded like a duet, and send activation to "duet", as well as to each other.)

(3) **Simplicity:** "The plausibility of a proposition is inversely related to the number of explaining statements needed to explain it." The simpler the explanation, the more likely it will be believed. That is, lots of assumptions (or co-beliefs) are often counterproductive, compared to fewer assumptions.

(4) **Data Priority:** "Results of observations have an extra measure (boost) of acceptability." This means that acknowledged facts, memories, and observations carry more importance than "mere" hypotheses.

(5) **Contradiction:** "Contradictory hypotheses incohere." This means that beliefs that conflict with each other send "negative activation" (or "inhibition") to each other, like rival members of two different "gangs."

(6) **Competition:** "Competing beliefs (which explain the same evidence or hypotheses but are not themselves explanatorily related) incohere." This means that highly independent explainers of the same proposition conflict with each other, and hence send "negative activation" to each other, like rival gang members vying for the same turf. (E.g., If you hear a report that an evil dictator was shot, and later hear that he was stabbed, you might assume that

the two reports offer competing hypotheses.) This principle may be optionally invoked in a variant of ECHO, called ECHO2, which automatically infers inhibitory relationships between propositions that independently (i.e., not *jointly*, as in principle 2) explain a third proposition.

(7) **Acceptability:** "The acceptability of a proposition increases as it coheres more with other acceptable propositions, and *incoheres* more with *unacceptable* propositions." This basically says that how much a belief is believed is a function of who its friends and enemies are, and how much *they* are believed.

(8) **Overall Coherence:** "The overall coherence of a network of propositions depends on the local pairwise cohering of its propositions." This basically means that the goodness of a whole "neighborhood system" of beliefs is determined by the believability of its members and their relationships.