"I Know One When I See One": How (Much) Do Hypotheses Differ from Evidence?

Michael Ranney
Patricia Schank
Christopher Hoadley
and
Jonathan Neff
Education in Mathematics, Science and Technology
Graduate School of Education
University of California, Berkeley, CA 94720

Institute of Cognitive Studies
Building T-4
University of California, Berkeley, CA 94720
ranney@cogsci.berkeley.edu
schank@garnet.berkeley.edu
tophe@garnet.berkeley.edu
jneff@uclink.berkeley.edu

Many discussions of (e.g., "scientific") reasoning highlight the classification of propositions into hypotheses (or "theory") and pieces of evidence (or "data"). Distinguishing between these belief categories is portrayed as a crucial prerequisite in developing critical thinking. In studying the principles of explanatory coherence and assessing our "reasoner's workbench" (*Convince Me)*, we find that the hypothesis/evidence distinction is nonobvious to undergraduates ("novices"). For instance, for isolated propositions, only zero-to-moderate (absolute) correlations were observed among "evidence-likeness," "hypothesis-likeness," and believability ratings. Inter-rater reliabilities for the evidence and hypothesis notions were also fairly low. Providing textual context for the beliefs, though, generally yielded higher discriminability and reliability (regarding hypothesis and evidence) across subjects — even for experts who study reasoning professionally. While training novices with *Convince Me* and its curriculum lends sophistication to these subjects' discriminative criteria, crisp classificatory delineations between hypotheses and evidence are still absent. Most disconcerting are findings that even experts have surprisingly low agreement when rating (a) a proposition's hypothesis-likeness or evidence-likeness, and (b) the goodness of novices' definitions of "evidence" and "hypothesis." These results are discussed regarding explanatory coherence, the associated ECHO model, *Convince Me,* and subjects' believability ratings. In sum, hypotheses and evidence may be closer to each other than we generally care to admit.

INTRODUCTION

How is it that context, as well as one's training, experience, and knowledge, affect the act of classifying an entity? Answering this question, as well as examining intra- and inter-classifier consistencies, are crucial — for one example — to assessing the stability of indexing in searches through databases and catalogs. We approach these topics from the "cognitive science of science" perspective, by examining the nature of essential categorizations in critical inquiry — in particular, the classification of statements as either evidence or hypotheses.

The distinction between evidence and hypothesis (or theory) appears to be a fundamental one in scientific reasoning (e.g., Kuhn, 1989). Researchers commonly suggest that it is desirable for children, and lay people (and perhaps even some scientists) to improve their understanding of this distinction. Our own simulations with the ECHO model (e.g., Ranney, Schank, Mosmann, &

ISSN: 2324-9773

Ranney, M. (1994). "I Know One When I See One": How (Much) Do Hypotheses Differ from Evidence?. 5th ASIS SIG/CR Classification Research Workshop, 139-156. doi:10.7152/acro.v5i1.13783

Montoya, 1993; Schank & Ranney, 1991), based upon the principles of the Theory of Explanatory Coherence (TEC; e.g., Ranney & Thagard, 1988; Thagard, 1989, 1992) also rely on the evidence/hypothesis contrast (as do other studies that have assessed principles of explanatory coherence, cf. Read & Marcus-Newhall, 1993). For instance, 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. 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. As Hanson (1958/1965) and other philosophers have pointed out, though, the observation/theory classification may not be clear-cut. Toward the end of his book, Hanson wrote: "...we have tried to explore the geography of some dimly-lit passages along which physicists have moved from surprising, anomalous data to a theory which might explain those data. We have discussed obstacles which litter these passages. They are rarely of a direct observational or experimental variety, but always reveal conceptual elements. The observations and the experiments are infused with the concepts; they are loaded with the theories." One might compare such an approach to Popper's (1978) whose answer to the chicken-and-egg question of "Which came first, evidence or hypothesis?" was: "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).

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 our "reasoner's workbench" software, *Convince Me*, 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.). We consider several questions regarding the hypothesis/evidence distinction: First, how are particular individual propositions classified? This is addressed by observing how subjects 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, we provide statements 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? In addressing this question, we compare samples of the two populations of subjects, focusing on both 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, we asked our novices to define the set of terms, and asked our experts to grade the goodness/accuracy of these novices' definitions.

We 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 TEC's data priority should lend activation (ECHO's currency of believability) 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).

METHOD

Subjects

Twenty subjects participated in this study. Ten subjects were undergraduate novices (four women and six men) from the University of California, Berkeley. 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 ten expert subjects were volunteers from the University of Chicago, the University of California (Berkeley), Princeton University, the Tennessee Institute of Technology, and the Educational Testing Service. (Five subjects were post-Ph.D., and five were doctoral students; three subjects 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 General Procedure

As shown in Figure 1, the novices completed a pre-test, three curriculum units on scientific reasoning, integrative exercises using the *Convince Me* software (for generating and analyzing arguments; Ranney, in press; Ranney et al., 1993; Schank & Ranney, 1993), and a post-test (which replaces some pre-test items with new isomorphic items). One subgroup of ("propositional") experts was asked to complete the proposition-rating portions of the pre-test (see below). The other subgroup of ("definitional") experts was given a randomly-ordered booklet of novices' completed

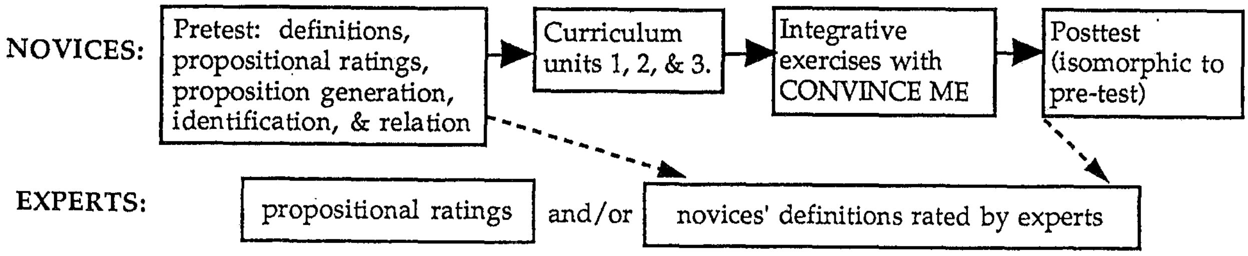

NOVICES:
Pretest: definitions, propositional ratings, proposition generation, identification, & relation
Curriculum units 1, 2, & 3.
Integrative exercises with CONVINCE ME
Posttest (isomorphic to pre-test)
EXPERTS:
propositional ratings
and/or
novices' definitions rated by experts

Figure 1. Summary of this experiment's method.

ISSN: 2324-9773

definitions from the pre- and post-tests (see (a) below), and were asked to score them on a scale from 1 (poor) to 3 (good) for each given definition.

Materials, Apparatus, and Specific Procedures

Pre-test (approximately 90 minutes). The pre-test assesses one's ability to classify hypotheses and evidence, evaluate scientific theories, and generate and disconfirm alternate hypotheses (using tasks that include some related to those of Wason, 1968, and Wason & Johnson-Laird, 1972). Subjects are further asked to (a) define hypothesis, evidence, fact, explanation, contradiction, theory, argument, confirmation bias, disconfirmation, recency bias, and primacy bias, (b) rate statements presented in isolation (see Table 1) or within a story context (see Table 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 would search for prototypical hypotheses and evidence; cf. Rosch, 1977),

Table 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) All wine is made from grapes.

definitely not hypothesis				neutral				definitely hypothesis
1	2	3	4	5	6	7	8	9

definitely not 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 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.

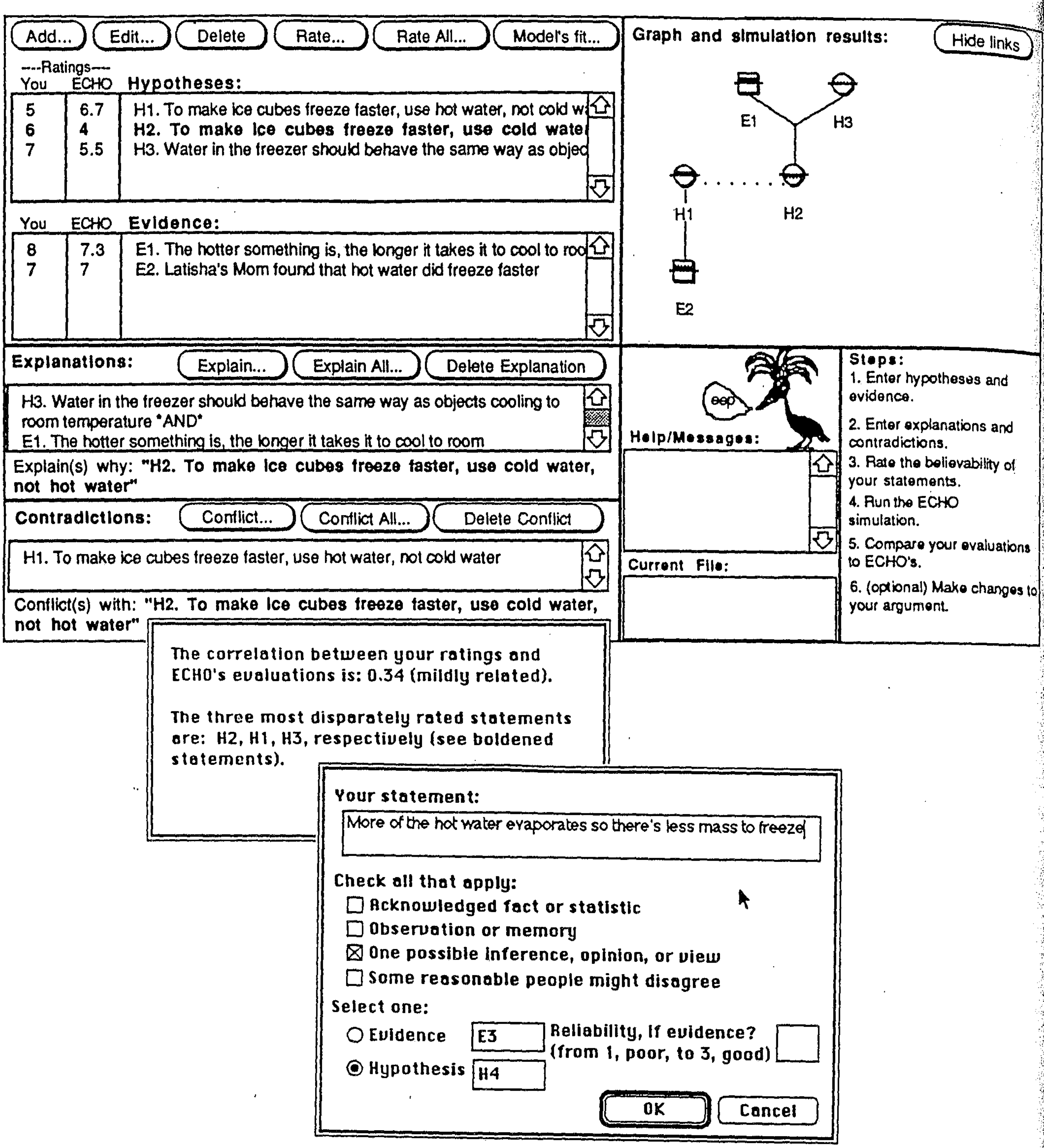
Examples of propositions from the above context, made available for rating:

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.

(c) generate hypotheses, attempt disconfirmations, and offer data regarding given situations, and (d) identify (and give believability ratings for) hypotheses and evidence in a given passage, propose (and rate) alternative propositions, state which propositions explain and contradict which others, and make any revisions to their argument and ratings.

Curriculum units (approximately one hour each, three hours total). Unit 1, "Evidence, Hypotheses, and Theories," is designed to help students (a) classify hypotheses and evidence, (b) generate and relate these to create "explanatory theories" — including justifications, and (c) evaluate the believability of the considered propositions. Unit 2, "Reasoning About Arguments," is primarily designed to help students generate alternative hypotheses and attempt to disconfirm them, all while reducing common biases (e.g., as observed in Ranney et al., 1993). Unit 3, "Using *Convince Me*," describes how to use this software to enter, save, and evaluate arguments. In particular, it explains to the student 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 to predict which beliefs "should" be accepted or rejected, based on their argument's structure, (6) contrast their ratings with ECHO's predictions, and (7) modify ECHO's parameters to better model their individual reasoning style, if they so desire. Throughout the curriculum, subjects are also encouraged to modify their arguments or ratings as needed.

ISSN: 2324-9773



---Ratings---

You	ECHO	Hypotheses:
5	6.7	H1. To make ice cubes freeze faster, use hot water, not cold w
6	4	**H2. To make ice cubes freeze faster, use cold wate**
7	5.5	H3. Water in the freezer should behave the same way as objec

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

Explanations:

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

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

Contradictions:

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/Messages:

Current File:

Steps:
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 to your argument.

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

Figure 2. A subject adds and classifies a belief about the speeds at which water of different initial temperatures freezes (bottom right) in response to *Convince Me*'s feedback (bottom left). (Note that subjects in this study used an earlier version of the software, which did not graphically display the argument structure—as shown in the upper right; rather, "activational thermometer" icons were merely displayed in rows and columns.)

ISSN: 2324-9773

***Convince Me* and integrative exercises** (approximately two hours). *Convince Me* elicits argument structures, each of which includes multiple theories, hypotheses, and pieces of evidence. After generating such an argument and offering believability ratings for one's elicited propositions, a student invokes *Convince Me*'s reasoning engine, ECHO, to obtain feedback on how well the student's structure seems to match his/her beliefs (as shown in Figure 2). Students may respond to this feedback by re-evaluating their beliefs, reformulating their arguments, or adjusting the ECHO model to better simulate their thinking (Ranney, Schank, & Diehl, in press). Subjects using *Convince Me* are given a set of integrative exercises, and are asked to both enter their arguments into the system and run the ECHO simulation. *Convince Me* runs using HyperCard (with external C commands) on a Macintosh with a 13" (Powerbook-size) or 17" (two-page) monitor (Schank, Ranney, & Hoadley, 1994).

Post-test (approximately 90 minutes). The 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). 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; (see (a) and (b) under "Pre-test" above). Four sets of items were isomorphic to those on the pre-test (involving hypothesis generation and disconfirmation, as well as the identification of evidence, hypotheses, explanations, and contradictions from a given passage, etc.; see (c) and (d) under "Pre-test").

Exit questionnaire (approximately 10 minutes). The exit questionnaire asks subjects 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. Subjects are given copies of the curriculum units to refer back to while completing the questionnaire.

RESULTS

As shown below, context and training generally improved subjects' hypothesis/evidence classifications, yet even experts found them difficult:

Propositional Ratings

Correlations among the constructs of evidence, hypothesis, and believability. As Table 3 illustrates, context generally adds to the discriminability between evidence and hypotheses across all types of subjects 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 (anti-)correlation. 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

ISSN: 2324-9773

Ranney, M. (1994). "I Know One When I See One": How (Much) Do Hypotheses Differ from Evidence?. 5th ASIS SIG/CR Classification Research Workshop, 139-156. doi:10.7152/acro.v5i1.13783

Table 3. Within-subjects correlations between believability and hypothesis-likeness (B-H), evidence-likeness and hypothesis-likeness (E-H), and believability and evidence-likeness (B-E).

	Novices			Experts		
	B-H	E-H	B-E	B-H	E-H	B-E
No context: pretest	.09	-.03	.48[a]	-.24[ab]	-.28[a]	.35[a]
post-test	-.14	-.30[a]	.60[a]			
In context: pretest	-.57[ac]	-.41[ac]	.42[a]	-.19[b]	-.66[abc]	.37[a]
post-test	-.68[ac]	-.63[abc]	.64[ab]			

[a] $r \neq 0$, $p<.05$; [b] significantly different from novice's pretest, $p<.05$; [c] significantly different from no-context, $p<.05$

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 subjects, context-types, and testing times. (Nb. that, after training, novices exhibited the largest of the believability-evidence correlations.)

Inter-rater agreement regarding the constructs of evidence, hypothesis, and believability. As shown in Table 4, both groups showed greater inter-rater reliability (correlations) across their

Table 4. Between-subjects correlations regarding believability (B-B), evidence-likeness (E-E), and hypothesis-likeness (H-H).

	Novices			Experts		
	B-B	E-E	H-H	B-B	E-E	H-H
No context: pretest	.66[a]	.31[a]	.15[a]	.87[ab]	.20[a]	.28[a]
post-test	.65[a]	.32[a]	.06			
In context: pretest	.20[ac]	.23[a]	.29[ac]	-.04[bc]	.42[abc]	.54[abc]
post-test	.25[ac]	.44[ab]	.39[ac]			

[a] $r \neq 0$, $p<.05$; [b] significantly different from novice's pretest, $p<.05$; [c] significantly different from no-context, $p<.05$

ISSN: 2324-9773

Ranney, M. (1994). "I Know One When I See One": How (Much) Do Hypotheses Differ from Evidence?. 5th ASIS SIG/CR Classification Research Workshop, 139-156. doi:10.7152/acro.v5i1.13783

believability ratings for the no-context propositions, regardless of testing time. This is not surprising, since the in-context situation (shown in Table 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 subjects 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 subjects' ratings of the propositions' believability. Ultimately consistent with this interpretation, pilot studies with experts showed that (a) assessing the present study's context-bound propositions out of context reduces 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.

Individual differences in ratings of the three constructs. Figure 3 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 subjects' 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 subjects, 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 subject rated (only) their *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

ISSN: 2324-9773

Ranney, M. (1994). "I Know One When I See One": How (Much) Do Hypotheses Differ from Evidence?. 5th ASIS SIG/CR Classification Research Workshop, 139-156. doi:10.7152/acro.v5i1.13783

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 subject (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 subjects 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 our 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.

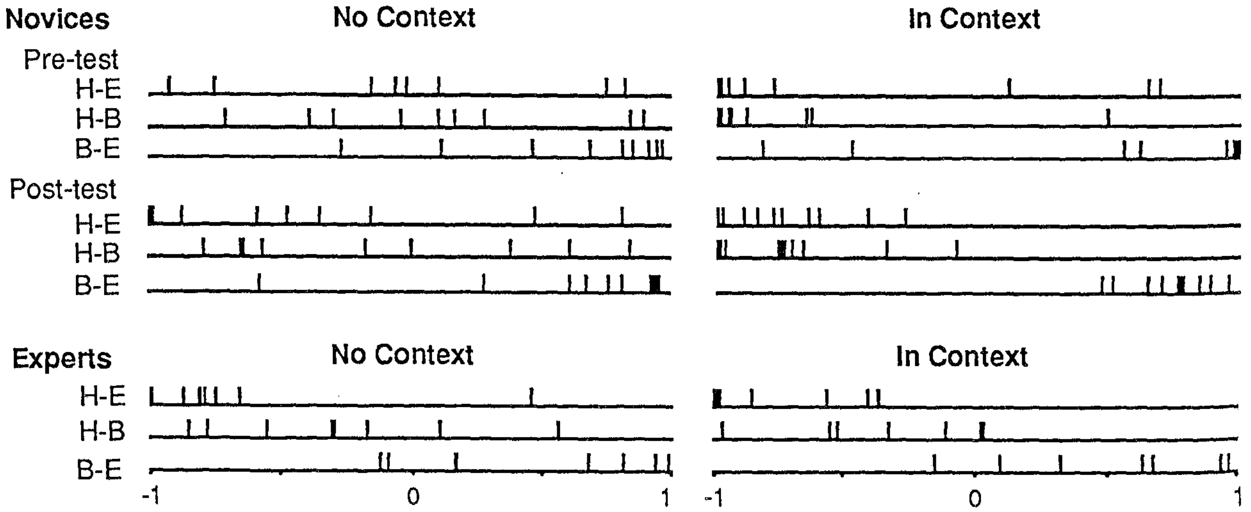


Figure 3. Novices' and experts' correlation distributions. (H-E, H-B, and B-E refer to hypothesis-evidence, hypothesis-believability, and believability-evidence correlations, respectively.)

ISSN: 2324-9773

Ranney, M. (1994). "I Know One When I See One": How (Much) Do Hypotheses Differ from Evidence?. 5th ASIS SIG/CR Classification Research Workshop, 139-156. doi:10.7152/acro.v5i1.13783

Table 5. Novices' mean pre-test definition scores, post-test change, and intercoder reliability correlations between five (expert) coders.

Definition	Pre-test mean score (3 points possible)	Mean changes from pre-test to post-test	Inter-rater reliability (between 5 coders)
hypotheses	2.10	+ 0.13	.39[a]
evidence	2.11	+ 0.16	.34[a]
fact	2.04	+ 0.44[b]	.14[a]
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[a]

[a] $r > 0, p<.05$; [b] significant increase from novice's pretest, $p<.05$;

Experts' Ratings of Novice's Definitions

Experts had considerable variety 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 for "contradiction" was not even statistically significantly above 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 only .21 for primacy bias). Table 5 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.)

DISCUSSION

Convince Me seems successful at improving novices' abilities to discriminate between the notions of evidence and hypothesis, as shown by their more negative correlations following training, as well as correlations involving believability that suggest that subjects take a largely positivistic approach. In accord with these findings, Table 6 presents some representative (albeit seemingly brochure-like) comments from the exit questionnaires of each of the ten subjects who used the curriculum with *Convince Me*.

The presence of a context also generally heightened the evidence-hypothesis distinction for our subjects (although experts seemed less positivistic across in-context propositions than the novices

ISSN: 2324-9773

Ranney, M. (1994). "I Know One When I See One": How (Much) Do Hypotheses Differ from Evidence?. 5th ASIS SIG/CR Classification Research Workshop, 139-156. doi:10.7152/acro.v5i1.13783

became). 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 contrast with views suggesting that belief evaluation is more likely to proceed top-down, with hypotheses recruiting either 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.

Table 6. A sample of subjects' comments about the system employing *Convince Me.*

Subject	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.*

That being said, it is perhaps most striking that our 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. Hence, we might contrast these findings with Supreme Court Justice Potter Stewart's pornography comment (from 1964): "I may not be able to define it, but I know it when I see it!" Regarding evidence and hypothesis, several grounds suggest that it is unlikely that people truly "know one when they see one." For one reason, our data 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 subjects decide upon ratings for a given statement. Furthermore, 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).

Table 7. Some factors that appear to influence a proposition's classification as a hypothesis or piece of evidence.

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

To appreciate the difficulty of categorizing even fairly straightforward propositions, consider one of our no-context stimuli, "President John F. Kennedy was assassinated." Many subjects see this as a piece of evidence. For them, the statement is essentially an observation, much like, "This rose is red." In contrast, our 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 truly believe that (a) the victim was Kennedy's double, (b) Kennedy survived the shooting, and/or (c) the event was an elaborate suicide. One can envision for nearly any statement a possible situation in which that statement is untrue or in doubt. So, for such theorists, the statement does appear hypothetical; for one to take it as a domain-independent fact is

Ranney, M. (1994). "I Know One When I See One": How (Much) Do Hypotheses Differ from Evidence?. 5th ASIS SIG/CR Classification Research Workshop, 139-156. doi:10.7152/acro.v5i1.13783

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 "assassinate"); on the contrary, the data collected so far suggest that these considerations are also those of our subjects. Indeed, these three divergent responses (i.e., evidence vs. fact vs. hypothesis) were recently independently elicited from (and respectively explained by) each of three Japanese cognitive scientists (who collaborate with one another on related topics) — for the very same statement. (Similar discussions with attorneys about these notions have yielded similar confusions.) 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, in concert with various findings from classification research.

Thus, apparent distinctions between evidence, facts, and hypotheses often appear more clear in the abstract than for concrete cases. "Theory-independent facts" often just mean "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." Indeed, this helps explain why our rating tasks, particularly for evidence and hypotheses, pose such difficulty. We can easily generate a dozen or more factors that influence the categorization of a proposition, but most of these involve aspects of context (see Table 7). Further, most of the context must be filled in by our subjects, even when the statement is embedded in a story. Contexts seem to merely reduce the space of scenarios in which a statement may serve, and this space often leaves many potential roles of both the evidential and hypothetical variety. In this respect, one either tries to mentally compare the relative likelihood of these two populations of roles, or one uses heuristics to try to find the proposition's most likely role. Such reasoning is similar to that proposed by Johnson-Laird's (1983) version of "mental models."

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

IMPLICATIONS AND FUTURE DIRECTIONS

These findings have considerable import for the field of science education. If we ourselves, as cognitive scientists, cannot readily and reliably label some notions as "evidence" and others as "hypothesis," how can we expect students to be terribly concerned about the distinction? Why should so many science textbooks start with descriptions of this distinction — or imply that the distinction is clear and apparent to experts? It would seem that there must be better — and more accurate — ways to both structure science instruction and portray the work of scientists.

Some have suggested that intensively training subjects (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, the sort of argumentation that we (and perhaps the

ISSN: 2324-9773

Ranney, M. (1994). "I Know One When I See One": How (Much) Do Hypotheses Differ from Evidence?. 5th ASIS SIG/CR Classification Research Workshop, 139-156. doi:10.7152/acro.v5i1.13783

classification research community) find more intriguing is that of "loose reasoning" (Ranney, in press), involving a complex of competing and supportive propositions that are generally neither absolutely implicative nor mutually exclusive in their relationships. Others have suggested that even less formal philosophical arguments, but those still largely dealing in the trade of premises and conclusions, may provide a more compelling model for the training of analyzers of arguments. 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 much of the contentiousness found in the literatures of such 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 classify propositions epistemically. Studies currently under way in our laboratory are addressing these questions, but both *a priori* analyses and preliminary evidence suggest that individual aptitudes and predilections will largely determine whether such graphics (e.g., of the sort shown in part of Figure 2) prove to be usefully synergistic representations or interfering rhetorical exercises. We can readily envision both potential benefits and dangerous biases arising from the use of either diagrammatic or textual/ statement-based representations of arguments (Diehl, Ranney, & Schank, in preparation). In other words, it is not yet clear — for the cognitive domain of evaluating controversial systems of beliefs — which representations will be most reasoning-congruent for which subpopulations of people (cf. reasoning-congruence and various representations in the depiction of computer programs; e.g., Ranney & Reiser, 1989; Reiser, Ranney, Lovett, & Kimberg, 1989; Reiser, Copen, Ranney, Hamid, & Kimberg, 1994).

CONCLUSIONS

Returning to our earlier empirical motivations, it appears that classifications are clearly modulated by the context of the to-be-classified entity, as well as the experiences and knowledge of the classifier. Training, as reified in our system, is also capable of significantly and rapidly moving novices toward the ranks of expert classifiers. The above measures of intra-classifier and inter-classifier concordance reinforce these generalizations.

More specifically, this study seriously (and empirically) questions the common implication that classifying evidence and hypotheses is reasonably straightforward once one formally studies science. On the contrary, it seems that even experts in scientific reasoning — including those who have studied the distinction themselves — have difficulty articulating data-theory differences. 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." In sum, we may not reliably "know one when we see one."

Convince Me and its associated paradigm improved novices' relatively deficient ability to discriminate between evidence and hypotheses, beyond the benefits of contextual embeddedness. This would appear to be an encouraging sign for developers of systems of this sort (cf. Cavalli-

ISSN: 2324-9773

Sforza, Lesgold, & Weiner, 1992). We may not be able to successfully give students pat definitions of complex epistemic concepts like theory and evidence, but we might aid their development through more sophisticated epistemological stages (e.g., Chandler, 1987). In the above, we see that, although even experts disagree on the distinction between theory and data, our reasoner's workbench certainly makes novices respond more like experts during such epistemic categorizations.

ACKNOWLEDGMENTS

We especially thank Christine Diehl and Kathryn Ann Dougery for their assistance with the design and implementation of the *Convince Me* curriculum. We also thank Brian Foley, Irvin Katz, Elijah Millgram, Michelle Million, Jeffrey Schank, Annie Selden, Giyoo Hatano, Kayoko Inagaki, Yuji Itoh, Hiroaki Suzuki, John Wesolowski, and Liz Wesolowski for their assistance and comments. Finally, we are grateful to the National Academy of Education, the Spencer Foundation, and the University of California's Committee on Research for supporting this study.

REFERENCES

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.

Chandler, M. (1987) The Othello Effect: Essay on the emergence and eclipse of skeptical doubt. *Human Development*, 30, 137-159.

Diehl, C., Ranney, M., & Schank, P. (in preparation). *Multiple Representations for Improving Scientific Thinking*. University of California, Berkeley.

Giere, R.N. (1991). *Understanding scientific reasoning*. New York: Holt, Rinehart, and Winston.

Hanson, N.R. (1965). *Patterns of discovery: An inquiry into the conceptual foundations of science.* London: Cambridge University Press. (Original work published in 1958)

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, 96, 674-689.

Mangan, B., & Palmer, S. (1989). New science for old. *Behavioral and Brain Sciences*, 12, 480-482. [Commentary on P.R. Thagard's "Explanatory coherence."]

Popper, K.R. (1978). *Conjectures and refutations* (rev. ed.). London: Routledge & K. Paul.

Ranney, M. (in press). Explorations in explanatory coherence. To appear in E. Bar-On, B. Eylon, and Z. Schertz (Eds.), *Designing intelligent learning environments: From cognitive analysis to computer implementation.* Norwood, NJ: Ablex.

Ranney, M. (1994a). Relative consistency and subjects' "theories" in domains such as naive physics: Common research difficulties illustrated by Cooke & Breedin. *Memory & Cognition*, 22, 494-502.

Ranney, M. (1994b). *Individual-centered vs. model-centered approaches to consistency: A dimension for considering human rationality.* Manuscript submitted for publication.

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* (pp. 88-95). New York: Elsevier Science Publishers.

ISSN: 2324-9773

Ranney, M. (1994). "I Know One When I See One": How (Much) Do Hypotheses Differ from Evidence?. 5th ASIS SIG/CR Classification Research Workshop, 139-156. doi:10.7152/acro.v5i1.13783

Ranney, M., Schank, P., & Diehl, C. (in press). Reducing the competence/performance gap with *Convince Me*, the reasoner's workbench. *Computers in Psychology Handbook.*

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* (pp. 101-106).

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

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. (1994). *Cognitive and Motivational Consequences of Tutoring and Discovery Learning*. Manuscript submitted for publication.

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* (pp. 228-235). Springfield, VA: IOS.

Rosch, E. (1977). Human categorization. In N. Warren (Ed.), *Studies in cross-cultural psychology* (Vol. 1, pp. 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?* (pp. 73-86). Hillsdale, NJ: Erlbaum.

Schank, P., & Ranney, M. (1991). An empirical investigation of the psychological fidelity of ECHO: Modeling an experimental study of explanatory coherence. In *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. In *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., & Hoadley, C.M. (1994). *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.

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

Thagard, P.R. (1992). *Conceptual Revolutions*. Princeton, NJ: Princeton University Press.

Tweney, R.D., Doherty, M.E., & Mynatt, C.R. (Eds.). (1981). *On scientific thinking*. New York: Columbia University Press.

Wason, P.C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology,* 20, 273-281.

Wason, P.C., & Johnson-Laird, P.N. (1972). *Psychology of reasoning: Structure and content.* Cambridge, MA: Harvard University Press.

ISSN: 2324-9773

ISSN: 2324-9773