Explanation and Evidence in Informal Reasoning

Sarah Brem
Department of Psychology
Northwestern University
2029 Sheridan Road
Evanston, IL 60208
sbrem@nwu.edu

Lance J. Rips
Department of Psychology
Northwestern University
2029 Sheridan Road
Evanston, IL 60208
rips@nwu.edu

Abstract

Explanation and evidence play important and non-interchangeable roles in argument. However, previous research has shown that subjects often confuse explanation and evidence (Kuhn, 1991). This study investigates the circumstances under which this confusion occurs. In Experiment 1, subjects generated arguments about issues of popular interest such as problems in schools and drug abuse. In Experiments 2 and 3, subjects rated the strength of evidence presented to them. The results of the protocol analyses and ratings tasks suggest that subjects tend to overestimate the strength of explanations when they lack sufficient knowledge of the domain or when they are unable to generate alternatives to the hypotheses presented to them. We consider reasons why relying on explanations in these circumstances might be a valuable heuristic.

Explanation and physical evidence both play important roles in arguments, but these roles are distinct and non-interchangeable. In many cases, explanations are a sort of causal story, while evidence informs us as to whether or not these stories are likely to be true. In creating and evaluating arguments, we need to keep these roles clear or risk erroneous conclusions.

In her book *The Skills of Argument*, Kuhn (1991) examined everyday reasoning on social issues. She found that subjects have difficulty producing multiple hypotheses and provide weak evidence to support their opinions. One of the more striking findings was subjects' inability to produce what Kuhn terms "genuine evidence," instead producing "pseudoevidence," even when arguing about familiar issues (Kuhn, 1991, Chapter 3). The present study investigates factors that contribute to subjects' difficulties in producing genuine evidence.

The distinction between genuine evidence and pseudoevidence is based on whether subjects produce a merely plausible tale of cause-and-effect, or whether they also provide evidence that the proposed cause actually does occur. One way to think of this is as a failure to distinguish between explanation and evidence. Both explanation and evidence must be taken into account in evaluating an argument; explanations provide causal mechanisms and motivate experiments and observations. But when creating an argument, it is important to be clear about which aspects of that mechanism have been substantiated, and which aspects are still based on unsubstantiated claims and assumptions. To the extent that any explanation relies on assumption, it should be viewed with a certain amount of caution. However, the subjects in Kuhn's study did not exhibit this caution.

In Kuhn's study, subjects were asked to give a reason why children fail in school and to say how they would prove they were right. If a subject explains how a parent's lack of interest could lead to their child failing in school, this explanation alone counts as pseudoevidence. If the subject supports this opinion by comparing children whose parents are highly involved with children whose parents are less involved, this would be genuine evidence. Kuhn further divided genuine evidence and pseudoevidence into subcategories, which we list in Table 1. Although it is possible to question some of these distinctions, we accept them temporarily in order to compare our own results to Kuhn's.

When Kuhn asked subjects to give reasons for everyday social problems, such as failure in school, subjects produced relatively few examples of genuine evidence. Although subjects claimed to be very familiar with the school failure topic, only 66% of college-educated subjects and 29% of subjects without a college education provided genuine evidence. Results for less familiar topics showed a similar pattern, although the amount of genuine evidence decreased across all groups.

In a subsequent phase, the same topics were presented a second time, but factual information was provided. There were two tasks. One involved underdetermined evidence. Several potential causes were presented, but the evidence was insufficient to draw any conclusion. The other involved overdetermined evidence: Experts gave evidence strongly supporting multiple causes. In both tasks, subjects tended to claim that the passage still supported their original opinion and, furthermore, that it did not support any alternative hypotheses.

Taken together, the results of Kuhn's study suggest that even people with academic training have difficulty providing and recognizing sound evidence. There appear to be at least two possible reasons for this: (a) Subjects possessed insufficient information to make their case, or (b) The facts were available, but subjects failed to distinguish genuine evidence from pseudoevidence.

Lack of relevant information. If subjects have no first-hand knowledge of an issue, they may have to settle for providing a plausible story. In Kuhn's protocol task, the wording may have led them to believe that they were limited

Table 1. Criteria for classifying responses, and examples of subjects' responses.

PSEUDOEVIDENCE

GENERALIZED SCRIPT: Subject explains how the proposed cause could bring about the effect without showing that the cause described actually occurs.

Why do African-Americans face greater economic hardship than Caucasians?

Because African-Americans generally come from a poor economic background, they are not given the same opportunities to develop as Caucasians. Thus, they have to struggle harder....

SPECIALIZED SCRIPT: Like a generalized script, but formulated as a specific example.

What causes apathy in teachers?

My mother was a schoolteacher who quit because she worked around the clock with students who didn't care and parents who didn't care for barely enough money to survive on.

GENUINE EVIDENCE

DISCOUNTING: Supports the proposed cause by undercutting rival causes.

What causes homelessness?

People are usually not born into this state of homelessness. I would show that many homeless people just fell on bad economic times and were forced out.

ANALOGY: Produces information about a second domain and shows that it is similar to the argument domain.

Why are the children of alcoholics likely to become alcoholics?

Children learn to speak...from watching their parents. If a parent has an accent it is probable that the child will develop this accent also....So if a child has an alcoholic parent...then they too might become alcoholics.

CORRELATION: The proposed cause co-occurs with the effect.

Why do criminals return to crime?

Check the number of criminals who return to crime who are brought up in a hostile environment or are from broken families.

COVARIATION: The effect is present when the cause is present, absent when the cause is absent.

What causes drug abuse among teens?

[Look at] reports showing the difference between teens who have jobs or extra-cirricular activities vs. those who don't and the degree to which they abuse drugs.

OTHER

NO RESPONSE. Question is left unanswered.

AUTHORITY. Subject stated that he/she would read newspapers or journals, or interview experts. The subject did not state what they would look for, or what they expected to find.

NON-EVIDENCE. A response is categorized as non-evidence when the subject:

1. Claims evidence is unnecessary, that the correctness of their opinion is self-evident.
2. Gives evidence establishing the effect. (E.g., when asked why failure in school occurs, subjects give evidence that failure does occur without stating why.)
3. Claims the effect does not exist. (E.g., subject claims school failure doesn't occur.)

to evidence they actually possessed. In the example of failure from school, subjects would have provided genuine evidence if they described a hypothetical study comparing involved to uninvolved parents. Unless they recognized that hypothetical evidence was legitimate, however, they would have to produce evidence they had previously encountered--a difficult task for non-experts.

Admittedly, the comprehension tasks do suggest that even when subjects have the relevant facts they cannot use them properly. However, the initial protocol task could have

influenced their later responses. Having already taken a stand on the issues could have: (1) pressured subjects to maintain consistency with their previous responses, or (2) hampered their ability to generate additional hypotheses (Koriat, Lichtenstein, & Fischhoff, 1980; Holt & Watts, 1969).

Failure to distinguish different kinds of evidence. A second explanation for Kuhn's results is that subjects were simply unable to distinguish genuine evidence from pseudoevidence. A related possibility is that subjects may have provided pseudoevidence because their criteria for a good argument were different from those of the experimenter. Science requires that we rule out alternative hypotheses, but less stringent criteria may be appropriate in everyday situations. We often have neither the time nor the resources to test all possible hypotheses, and some factors are beyond our control, such as replacing the teacher when your child does poorly in school. What passes as a good argument in practical terms may not be a good *scientific* argument.

Having identified some possible causes of the results described above, we now turn to the experiments that may help us to differentiate them.

Experiment 1: Can lack of knowledge account for failure to produce evidence?

The first experiment tested the possibility that the subjects in Kuhn's experiments may have suffered from a lack of relevant facts, coupled with a failure to understand the hypothetical demands of the task. The procedure was similar to that used in Kuhn's protocol task, except that subjects were assigned to one of two conditions: Ideal or Actual. In the Actual condition, we solicited evidence using the same wording as in Kuhn's study. In the Ideal condition, we asked subjects to give the strongest supporting evidence they could imagine. If subjects in Kuhn's study provided explanations to make up for their lack of sufficient information, they should produce more genuine evidence in the Ideal condition than in the Actual condition.

Method

We asked subjects about their opinions on 16 different issues (e.g., why do children fail in school?). Subjects were asked to write down their opinion on an issue and then rate their familiarity with that issue. Familiarity ratings were made on a 0 to 7 scale. Next subjects were asked to provide one or two pieces of evidence to support their opinion. Subjects in the Actual condition were asked "If you were trying to convince someone your view is right, what evidence would you give to try to show this?" Subjects in the Ideal condition were asked, "If you were trying to convince some else that your view is right, what would be the ideal evidence to show this? Imagine that you have access to any information or techniques you require." Finally, all subjects were asked to rate the strength of their own evidence on a 0 to 7 scale. All questions pertaining to an issue were presented on the same page of a booklet.

Twenty paid subjects participated, 10 in the Actual condition and 10 in the Ideal condition. All had completed at least two years of college and were native speakers of English.

Results

The results were scored by two raters, one who was blind to the hypothesis. Disagreements were resolved through discussion. The raters categorized the subjects' responses as genuine evidence, pseudoevidence, or other. These categories were further broken down into the sub-categories used by Kuhn (1991). The criteria for classification and examples of subjects' responses are given in Table 1. For the purposes of analysis, responses categorized as other were excluded, as they cannot be clearly incorporated as genuine evidence or pseudoevidence.

In the Actual condition, 34.8% of responses were classified as genuine evidence; 57.6% of responses in the Ideal condition were genuine evidence. The difference between conditions is significant (by subject, $t(18) = 2.25$, $p < .05$; by item $t(30) = 3.51$, $p < .001$). As shown in Table 2, Actual and Ideal differ primarily in three subcategories: correlational instances, generalized scripts, and references to authoritative sources. Subjects in the Actual condition produced 75 generalized scripts (i.e., general descriptions of possible cause-effect relations), 24 instances of correlational evidence, and 6 references to authority. Subjects in the Ideal condition produced 41 generalized scripts and 43 instances of correlational evidence, and 35 references to authority.

There was little difference between the familiarity and satisfaction ratings in the two conditions. Subjects in the Actual condition gave a mean familiarity rating of 3.82 (SD

Table 2. Frequency of evidence types by sub-category (Experiment 1).

	ACTUAL	IDEAL
GENUINE EVIDENCE		
Correlation	24	43
Covariation	13	18
Correlated Change	2	5
Analogy	1	0
Discounting	6	2
TOTAL	46	68
PSEUDOEVIDENCE		
Generalized Scripts	75	41
Specialized Scripts	11	9
TOTAL	86	50
OTHER		
Authority	6	35
Non-evidence	6	4
No Response	16	4
TOTAL	28	43

= 1.19), and a satisfaction rating of 3.46 (SD = 1.25). In the Ideal condition, subjects gave a mean familiarity rating of 3.61 (SD = 1.00), and a satisfaction rating of 3.90 (SD = 1.06). T-tests showed that the differences between conditions were not significant (Familiarity: $t(18) = 0.44$, $p > .10$; Satisfaction: $t(18) = 0.85$, $p > .10$).

Discussion

The results of Experiment 1 suggest that subjects' inability to provide appropriate evidence is at least in part due to a lack of relevant facts. The topics addressed in both Kuhn's studies and this experiment are matters of popular interest, but they are complex issues, and providing a sound analysis involves a considerable amount of data and effort. Unless subjects rely on hypothetical data, they are not likely to succeed at the task. Subjects in the Actual condition felt constrained to rely only on their own current knowledge, which was generally too meager to support genuine evidence. Subjects in the Ideal condition, however, were able to invent stronger evidence or seek an appropriate source of information. This implies that many subjects have an understanding of what makes good evidence and are able to make a strong case under favorable circumstances.

Although subjects performed considerably better in the Ideal condition, 31% of responses were scripts, and 82% of these were generalized scripts, which provide no evidence whatsoever that the proposed cause exists. Also, subjects' satisfaction with these generalized scripts was roughly the same in both conditions (Actual condition, 3.81; Ideal condition, 3.79). This suggests that, rather than being aware of producing inferior evidence, subjects in the Actual condition were as content with their result as subjects in the Ideal condition. It may be that subjects are drawn to scripts even when they are not limited by their personal resources. We wished to determine whether subjects would still consider scripts good evidence when they did not produce the evidence themselves.

Experiment 2: Can people evaluate evidence accurately?

Experiment 2 was conducted to determine whether the subjects would perform the same when presented with evidence as did the subjects in Experiment 1, who produced their own evidence. If the earlier subjects produced scripts simply because they were unable to come up with better evidence, they should nevertheless be able to recognize genuine evidence.

Method

Stimuli consisted of 16 sets of stories. The issues were the same as those in Experiment 1. For each issue, we generated an opinion and evidence supporting that opinion. There were eight types of evidence: generalized script, specialized script, discounting, covariation n = 1 (single instance), covariation n > 1 (multiple instances), correlated change n = 1, correlated change n > 1, and field study (see Table 1). The types of evidence are distinguished by the same criteria used in Experiment 1, except for field studies, which were not produced in Experiment 1. Field studies are defined as having an authoritative source, random assignment, and statistically significant results.

Additionally, 16 filler items were included. Each filler item consisted of an opinion and a piece of evidence concerning some issue, and two filler items were generated for each of the eight levels of evidence. Fillers were indistinguishable from the test items.

Subjects were presented with the 16 test items and 16 filler items. For each issue, subjects saw only one piece of evidence, and all items were presented on a separate page. Subjects were asked to rate the strength of each piece of evidence on a 7-point scale, 7 indicating the strongest evidence. After a 15 minute interval, subjects completed a recall task involving the rated items. The recall task is not relevant to this study and will not be discussed further.

Twenty-four Northwestern undergraduates received class credit for their participation. All were native speakers of English.

Results

The mean strength ratings are presented in Table 3. There was a significant difference between subcategories in both a by-item analysis ($F(7,105) = 5.46$, $p < .01$) and a by-subject analysis ($F(7,154) = 7.59$, $p < .01$) The mean rating for genuine evidence (3.86) was *lower* than for pseudoevidence (4.13). A planned comparison contrasting scripts with genuine evidence showed a marginally significant difference (by subjects: $F(1, 154) = 2.85$, $p < .10$; by item: $F(1,105) = 3.38$, $p < .10$).

Discussion

Subjects showed a trend toward preferring scripts over genuine evidence. This is consistent with Kuhn's finding; subjects seem willing to accept scripts as good evidence. One reason subjects may not have distinguished between explanations and evidence is that they saw only one piece of evidence for each question, and were unable for this reason to pick out what is most important about the supporting evidence. Comparing two items can highlight differences, allowing subjects to reflect on which elements are significant (Gentner & Markman, 1994). In essence, subjects were being placed in an impoverished condition by seeing only one piece of evidence supporting each opinion. Subjects who have insufficient information may place a higher value on scripts.

Experiment 3: Does context help subjects recognize genuine evidence?

The stimuli were the same as those in Experiment 2, with the omission of the filler items. Subjects saw all 16 question/opinion pairs and all eight levels of evidence for each question. The evidence supporting an opinion was presented together on a single page, and subjects rated the strength of each piece of evidence on a 0 to 7 scale (7 = strongest evidence). We encouraged subjects to compare all the evidence associated with an opinion before making their ratings.

Twenty Northwestern undergraduates participated, receiving class credit. All were native speakers of English.

Table 3. Mean ratings from 2 and 3, in which subjects rated the strength of presented evidence on a 0 to 7 scale (7=strongest).

PSEUDOEVIDENCE	EXPERIMENT 2	EXPERIMENT 3
Generalized Scripts	4.09	3.24
Specialized Scripts	4.25	2.96
GENUINE EVIDENCE		
Discounting	2.58	2.66
Correlation, N = 1	3.65	3.44
Covariation, N = 1	4.00	3.90
Correlation, N > 1	4.00	3.51
Covariation, N > 1	4.23	4.11
Field Study	4.35	4.66

Results

The mean strength ratings appear in Table 3. As in Experiment 2, the evidence subcategories differed in perceived strength ($F(7,133) = 7.90$, $p < .01$ by subjects, $F(7,105) = 33.67$, $p < .01$ by item). This time, ratings forgenuine evidence were higher, on average, than ratings for scripts (3.72 vs. 3.12). A planned contrast of scripts vs. genuine evidence showed a significant difference ($F(1,133)=14.90$ $p<.01$ by subjects, $F(1, 105)=44.47$, $p<.01$, by item).

We should note that the strength of genuine evidence in Experiments 2 and 3 may be reduced by the inclusion of discounting evidence (see Table 1 for an example of discounting). In both experiments, subjects gave unusually low ratings to discounting, as shown in Table 2. Kuhn classifies discounting as genuine evidence. However, while discounting argues against rival causes, it fails to show that the proposed cause actually occurs. In this way, discounting is not unlike scripts. If we remove discounting from our analysis, the difference between scripts and evidence in Experiment 2 decreases (Scripts: 4.13 vs. Evidence: 4.04), while in Experiment 3 the difference increases (Scripts: 3.12 vs. Evidence: 3.92). Therefore, providing subjects with a variety of evidence may have a greater beneficial effect than was initially suggested.

Discussion

In contrast to the results of Experiment 2, the results of Experiment 3 suggest that subjects can distinguish between evidence and scripts. This experiment differed from Experiment 2 in that subjects saw more supporting evidence for each opinion, and they were explicitly encouraged to compare different types of evidence. Subjects may have had difficulty generating alternative evidence in Experiment 2 and therefore failed to consider that better arguments could be made. When that evidence was provided for them in Experiment 3, their accuracy increased.

General Discussion

The purpose of this study was to determine whether subjects distinguish between explanations and evidence. Taken together, the results suggest that subjects do understand the difference, but that this distinction blurs when resources are limited. If subjects have strong evidence at their disposal, they will tend to recognize that explanations make for a fairly weak argument. When evidence is unavailable, subjects place more weight on explanations.

Available information may affect subjects in more than one way. In Experiment 1, we showed that subjects were capable of inventing evidence when encouraged to do so. In Experiment 2, however, subjects apparently failed to consider that there could be other arguments besides the one presented to them. When we gave them examples of these alternatives in Experiment 3, they may have recognized the merits of each and were therefore able to give a more accurate appraisal.

The ability to conceive of alternatives may play an important role in evaluating arguments, and encouraging subjects to generate alternatives may improve their argumentation skills. The importance of alternative hypotheses has been argued by Kuhn (1991). Subjects who did not generate alternative hypotheses were more confident and less realistic in evaluating their arguments than subjects who were able to imagine other possibilities.

The pattern that emerges from our data is that when subjects are faced with limited resources, they will tend to rely more heavily on explanations as support for a particular position. It is important to point out that this, in itself, is not necessarily an irrational or problematic strategy. A number of authors have pointed out that presumptive reasoning, that is, reasoning based on supposition and unsubstantiated claims, plays a vital role in informal reasoning. (e.g., Walton, 1992). There are many decisions to be made and debates to be resolved in environments in which the concrete examples and statistics are not available, and may never be available. In those cases, rather than shutting down altogether, people show great flexibility and

ingenuity by asking "What if..." and following that line of reasoning, with all of its assumptions, to its conclusion.

Presumptive reasoning has a number of valuable aspects: it can uncover internal inconsistencies in reasoning; it can make clearer what issues and questions are involved in a particular debate or decision (Walton, 1992); it may suggest new experiments or sources of information. Along similar lines, causal scripts can make a hypothesis more plausible by illustrating a logically possible route from the proposed cause to its effect. Although, as Kuhn (1991) points out, relying heavily on unsupported explanations can be a poor strategy, since plausibility does not ensure correctness, plausibility may be a good heuristic. Problems such as teacher apathy do not have a simple answer; we may never identify all the factors involved. And these issues are not only intellectual puzzles, but real problems, often requiring a quick, decisive response. The plausibility heuristic may narrow down the possibilities to a manageable number.

Pennington and Hastie (1986, 1992) have put forth a related argument, claiming that creating plausible stories is a spontaneous and important precursor to forming an opinion. They show that jurors organize evidence into a story before rendering a verdict. Their model posits that contructing a story helps subjects in a number of ways, including determining the significance of and finding the holes in their interpretation of a case. Although our subjects may not use evidence explicitly in constructing their stories, organizing their thoughts in this way may help them to determine the significance of possible factors and to maintain consistency.

Several other lines of research have suggested additional reasons why a reliance on explanation may be a productive strategy. For example, in the domain of electricity, Schauble, Glaser, Raghavan, and Reiner (1992) found that subjects' grasp of domain-general principles of evidence generation depended on their understanding of electricity. Without a solid theory of how circuits work, subjects could not determine what constituted a good experiment. Likewise, in everyday reasoning, if subjects do not have a good theory of a phenomenon, they may have difficulty identifying strong evidence. In such a situation, using scripts may seem like a good idea, in that they layout a causal model.

Also, subjects' interest in explanations seems quite sensible if sensitivity to causal mechanisms is judged important in causal attribution. We may be uncomfortable with unsubstantiated explanations of an event, but we also find observations without explanations unsatisfying. Ahn, Medin, Kalish, and Gelman (1995) found that subjects who were asked to determine the reason for a particular event focused much more heavily on data relating to possible causal mechanisms than to covariational data. Subjects apparently do not simply want to know that there is a statistical relationship between a cause and event, but they also want to know why.

The danger in presumptive reasoning is failing to recognize that one is reasoning from assumptions and unsubstantiated claims. If the world fails to behave as we predict, these are the links in our theory that should receive the largest share of our initial suspicions and come under the closest scrutiny. To the extent that unsubstantiated explanations are vulnerable in this way, it seems reasonable to view them as somewhat weaker than claims supported with evidence, and it is important is that we remember that these are merely assumptions when things go awry. An interesting follow-up to the studies described here would be to determine whether subjects do maintain an awareness of which claims are supported and which are not.

References

Ahn, W., et al. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, 54, 299-352.

Gentner, D., & Markman, A.B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science, 5*, 152-158.

Holt, R., & Watts, D. (1969). Salience of logical relationships among beliefs as a factor in persuasion. *Journal of Personality and Social Psychology, 11*, 193-203.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 107-118.

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

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology, 51*, 242-258.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story model for juror decision making. *Journal of Personality and Social Psychology, 62*, 189-206.

Schauble, L., Glaser, R., Raghavan, K., & Reiner, M.(1992). The integration of knowledge and experimentation strategies in understanding a physical system. *Applied Cognitive Psychology, 6*, 321-343.

Walton, D.N. (1992). *Plausible argument in everyday conversation*. Albany: State University of New York Press.