Explanations of the Use of Reliability Information as the Response in Probabilistic Inference Word Problems.

Robert M. Hamm, PhD
Institute of Cognitive Science
Box 345
University of Colorado,
Boulder, CO 80309

303/492-2936

December, 1987

Institute of Cognitive Science
Publication Number 87-13

This paper was presented at the Psychonomic Society meetings, Seattle, November, 1987. Comments from Maya Bar-Hillel, Gary Bradshaw, Reid Hastie, Richard John, Sarah Lichtenstein, Paul Smolensky, and Amos Tversky have stimulated changes and are gratefully acknowledged. Mitch Nathan helped with OPS5 simulations, and Janet Grassia helped with the figures. The work was sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences, Contract MDS-903-86-K-0265.

Table of Contents

Robert M. Hamm
December 3, 1987
Explanations for Reliability as Response.

1. Abstract.

Probabilistic inference word problems, such as the Blue/Green Cab problem, require subjects to state p(H/E), the probability of a hypothesis H, when they are given information about base rate p(H), evidence E, and the reliability of the evidence p(E/H). A frequent wrong answer is the reliability. Data from a study in which subjects answered after receiving each piece of information are used to evaluate three explanations for the use of reliability. Production system models state each hypothesis unambiguously. The hypothesis that subjects consider the baserate to be irrelevant in principle is rejected. The data are consistent with both the hypothesis that subjects confuse p(E/H) with p(H/E), and the hypothesis that they interpolate between the base rate probability and 1.0 but select their response from among nearby numbers that are available in the word problem.

Explanations of the Use of Reliability Information as the Response in Probabilistic Inference Word Problems.

2. Introduction.

It has been hypothesized that people use heuristic strategies to answer problems requiring statistical reasoning (Tversky and Kahneman, 1974; Kahneman, Slovic, and Tversky, 1982). It is assumed that a number of such strategies are available, that their use is contingent on the particulars of the situation, and that they usually produce answers that are approximately correct. Though this notion is appealing, Hastie (1983) has noted that neither the nature of the heuristic strategies nor the conditions in which they are used have been clearly articulated.

The present research follows Simon's (1976) suggestion that the techniques and theory of the information processing approach be used to specify the strategies, thus allowing them to be tested. It focusses on probabilistic inference word problems, where there is controversy about what strategy might responsible for a common error. Production system models are used to specify the competing heuristic strategy explanations (see also Johnson and Payne, 1985).

Probabilistic inference word problems (Bar-Hillel, 1980; Fischhoff and Bar-Hillel, 1984; Fischhoff, Slovic, and Lichtenstein, 1979; Kahneman and Tversky, 1972; Tversky and Kahneman, 1982), such as the Blue/Green Cab problem (Table 1), require subjects to estimate the probability that an explanation for an event is true, given two kinds of pertinent information: statistical or base rate information, and unreliable evidence about the case in question. Typically, the base rate and the evidence are in conflict: "Event A usually happens, but today the uncertain evidence points to B. What is the probability that B happened today?"

****

Insert Table 1 about here.

****

For example, the base rate p(H) information in the Blue/Green Cab problem is that 15% of the cabs in the city are Blue. The evidence E is that the witness identified the cab as Blue. The reliability of the evidence p(E/H) is that in a test 80% of the cabs were identified correctly. The required answer is the probability of the hypothesis given the evidence, p(H/E): the probability that the guilty cab was Blue given that the witness said so. When all three pieces of information are available, the correct answer is p(H/E) = .41, as calculated by Bayes' Theorem (see below).

The best established finding has been that when presented with base rate and case information, people neglect the base rate. This means that the probabilities they assign to the hypothesis are closer to 1.0, and farther from the base rate .15, than they ought to be (Bar-Hillel, 1980). A second finding is that people use available numbers. For example, the typical wrong answer when all three pieces of information are available is the reliability information, e.g., .80 in the Cab problem. Many other wrong answers are also available numbers, as are both wrong and right answers when only a subset of the information has been given (Hamm, 1987a).

The reliability response has attracted special attention, both because it is the most frequent response, and because hypotheses are available to account for it. This paper will focus on three hypotheses, each of which says that the reliability is used due to special characteristics it happens to have. The first hypothesis, due to Cohen (1981) and Niiniluoto (1981), is that subjects ignore the base rate on principle and apply normative procedures that produce the reliability as the correct answer. The second hypothesis, from Eddy (1982) and Dawes (1986), is that subjects confuse the symbolic representations of the conditional probabilities p(H/E) and p(E/H). A third hypothesis holds that subjects integrate base rate and case information but "neglect", i.e., give insufficient weight to, the base rate (Bar-Hillel, 1980, and Tversky and Kahneman, 1982); the result of their integration happens to be in the vicinity of the reliability.

3. Methods.

Subjects were 265 undergraduate students, 131 males, who participated for course credit in groups of from 15 to 40 subjects. They individually completed a questionnaire with 7 word problems, at their own pace. It was explained that although they probably have not been explicitly trained in the methods for solving these problems, there are indeed correct answers. Subjects were exhorted to pay serious attention to the problems, and were promised that their answers would be scored and the scores posted publicly.

Three of the seven problems were probabilistic inference word problems, occupying positions 3, 5, and 7 in the questionnaire. The four filler problems required the estimation or calculation of numerical quantities. The probabilistic inference word problems are the Cab problem used in previous research (Table 1), plus problems concerning the probability that a patient has a particular disease, and the probability that a particular twin boy broke a lamp (see Hamm, 1987a). Each problem was divided into four paragraphs, containing the introduction, the base rate information p(H), the evidence, and the reliability of the evidence p(E/H) and p(~E/~H). The subject was asked for the probability of the hypothesis (e.g., that the cab involved in the accident was a Blue cab) and its complement (that it was a Green cab) after each paragraph. Subjects were instructed to cover each page with a sheet of paper and slide it down to expose only one paragraph at a time (see Table 1). They were explicitly permitted to refer back to earlier paragraphs within a problem at any time. The base rate (b), evidence (e), and reliability (r) information within each problem were presented in each of the six possible orders for different subjects.

4. Results.

The method of stepwise presentation of the key information in probabilistic inference word problems allows observation of responses to all possible combinations of the information. The most popular answers on the Blue/Green Cab problem are shown in Table 2. The row labels indicate the available numbers .15 (the base rate), .50 (available a priori, indicating equal chance it was a Blue or a Green cab), .80 (the reliability), and 1.0 (available a priori, indicating complete belief that the guilty cab was Blue). The columns represent each of the possible information conditions, ranging from having no information to having all three pieces of information. Some answers are missing because subjects wrote verbal phrases or gave ranges. When none of the information had been presented, 252 of 256 subjects responded with .50. When only base rate information .15 had been presented, 40 of 83 subjects used it as their response. When baserate, evidence, and reliability had been presented (in any order), 101 of 247 subjects used the reliability .80 as their response, in accord with the findings of previous studies (e.g., Bar-Hillel, 1980).

****

Insert Table 2 about here.

****

The hypotheses explaining why the reliability information is used as the response will be evaluated by constructing production system models for each and comparing the predictions of these models with the data. A production system, as embodied in the OPS5 language (Brownson, Farrell, Kant, and Martin, 1985), is a collection of rules. Each rule specifies what the subject would do in a particular situation, i.e., when given a particular set of information. Collectively the rules cover all the externally defined situations in the study, plus all situations that can be produced internally by rule applications.

The production systems are used here to specify plausible psychological strategies and the situations in which the strategies would be applied. Therefore their rules make few assumptions about the details of information processing (STM, etc). Rather, the rules are couched in terms of the information that is available to the subject, the responses the subject makes, and judgment and memory operations the subject might apply to the available information. The operations are described in terms that are intended to specify them enough that the reader can judge whether it is plausible that a person would do that in such a situation. This plausibility is a feature on which the models may be evaluated.

After constructing a coherent and plausible set of rules that expresses the processes a hypothesis refers to, the hypothesis is evaluated by applying the rule system to each of the information conditions in the study to see whether it produces the most popular answers. If a particular production system predicts the typical subject's answers on the Cab problem, including the use of reliability when all information is

given, then the hypothesis it represents is a sufficient and plausible explanation for the use of reliability.

5. The Three Hypotheses and their Production System Models.

I will show first that the subjects' answers are inconsistent with the hypothesis that people universally ignore base rate on principle. Then I will show that both of the other hypotheses are consistent with the data.

5.1. Hypothesis 1: Principled Ignoring of Base Rate.

L. J. Cohen (1981) argued that on probabilistic inference word problems subjects might, with some justification, judge the statistical information to be irrelevant. For the Cab problem, their justification might be the principle that a defendant is "innocent until proven guilty," which implies that a statistically-based prejudice does not constitute proof. Other problems have their own particular forms of the argument that the base rate does not apply to the particular case, and subjects have been observed to articulate these principles (e.g., Lyon and Slovic, 1976).

Niiniluoto (1981) demonstrated further that if subjects do not believe that the relative frequency of Blue and Green Cabs in the city is pertinent to the question of which color of cab was involved in a particular traffic accident, then they would be correct to use the reliability .80 as their answer. Their reasoning could go as follows:

1. ignore base rate information as irrelevant to the prior probability p(H),
2. use a prior probability of p(H) = .50,
3. apply Bayes' Theorem, yielding p(H/E) = p(E/H)

$$p(H/E) = \frac{p(E/H)\times p(H)}{p(E/H)\times p(H) + (1-p(E/H))\times(1-p(H))}$$

$$= \frac{p(E/H)\times .50}{p(E/H)\times .50 + (1-p(E/H))\times .50}$$

$$= p(E/H)$$

4. and report the reliability, p(E/H), as the estimate of p(H/E).

5.1.1. Production system model.

The Principled Ignoring of Base Rate Hypothesis is embodied in the production system shown in Table 3. It has 4 rules (numbered arbitrarily to maintain consistency with rules in production systems presented below and in Hamm, 1987a and 1987b). Rule 1 asserts that if there is no pertinent information, the subject will respond with p(H/E) = .50. Rule 3 says that if there is evidence with no indication that it is unreliable, the subject will fully believe the hypothesis the evidence implicates. Rule 5 says that if the reliability of the evidence is known, then the subject will apply Bayes' Theorem. Rule 12 applies Bayes' Theorem with the assumption that the two hypotheses are equally likely.

****
Insert Table 3 about here.
****

Together these rules express the hypothesis, because there is no rule that takes account of the base rate information. Whenever a prior probability is used, it is the .50 of no information, rather than the base rate. Thus the behavior of this production system model is consistent with the notion that subjects universally ignore base rate information on principle, and that they otherwise act in accordance with the norm of Bayes' Theorem. Is it also consistent with the data?

Table 4 shows the rules that would be applied in each of the conditions of the study, and the

answers that would be produced by the application of these rules. For example, in the **no information** condition, Rule 1 is applied, producing an answer of .50. In the **b&e&r** condition, where base rate, evidence, and reliability are all given, Rules 5 and then 12 are applied, producing the answer .80. These predictions should be compared with the most common answers in Table 2.

****
Insert Table 4 about here.
****

The production system successfully predicts that .80 will be the most common answer in the condition where all three pieces of information are present, **b&e&r**. It also predicts this for the evidence and reliability **e&r** condition, which is a new and correct prediction (see Hamm, 1987a). However, to fully evaluate the Hypothesis of Principled Ignoring of Base Rate, we must look at the answers in all conditions. Of particular interest are the **no information**, reliability **r**, base rate **b**, and base rate and reliability **b&r** conditions.

In the **no information** and **r** conditions, the most common answer was .50 (see first and fourth columns in Table 2). This shows that people know how to use .50 as an expression of p(H) when there is no pertinent information, as the Principled Ignoring of Base Rate Hypothesis presumes they will also do when only base rate information is available.

In the **b** and **b&r** conditions, there is statistical (base rate) information, but no case information. If subjects consider the statistical information to be irrelevant, then they ought to ignore it and answer .50. We find, instead, that the base rate was the most popular answer. In fact, ten times as many subjects used the base rate as used .50 (see second and fifth columns in Table 2).

These results disprove the hypothesis that people universally ignore the base rate information. Note that similar results were found in the **b** condition by Kahneman and Tversky (1972).

A variant of the Principled Ignoring of Base Rate Hypothesis is that subjects would ignore base rate when there is any case evidence at all, but use it if it is the only information. Our data is consistent with this variant, and a production system could be constructed to express it by adding a rule:

```
Rule 2:
 If
   Query p(H)
   Base rate
   No evidence
 Then
   Answer = Base rate
```

However, this variant does not express the universal principle that statistical information is irrelevant, which is the essence of Cohen's (1981) argument. Further, the fact that no subjects gave any indication of applying Bayes' Theorem (no calculations in the margins in this questionnaire study; no verbal reports in think aloud studies) makes Rule 12 somewhat implausible. Thus it unlikely that this or any other variant of the Hypothesis of Principled Ignoring of Base Rate is true.

5.2. Hypothesis 2: Confusion of Conditional Probabilities.

The reason so many subjects answer with .80 when given all three pieces of information on the Blue/Green Cab problem may be that they have difficulty interpreting the conditional probability p(E/H), "the probability that the witness would say 'blue' if the cab were truly blue." They confuse this with the p(H/E) idea, "the probability that the cab was truly blue if the witness said it was 'blue'."

Eddy (1982) and Dawes (1986) have suggested that difficulty discriminating these concepts may be at the root of the base rate fallacy. Eddy, for example, found that even in their professional writings, medical doctors have confused p(cancer/positive test), the probability that a patient with a positive test

result has cancer, with p(positive test/cancer), the probability that a patient who has cancer will yield a positive test (see also Widiger et al, 1984). This is not to say that people can not understand each of these concepts. But when they must interpret symbolic expressions of the concepts, as when my subjects solve these word problems or when clinicians try to derive guidance for treating patients from summary statistics concerning the relation between symptoms and diseases, there may be some slippage between the conditional probability that the writer meant and what the reader understood (see Pollatsek et al, 1987, p 268).

Further evidence supporting the hypothesis comes from studies of the effects of training programs. Christensen-Szalanski and Beach (1982), Lichtenstein and MacGregor (1984), and Pollatsek et al (1987) have given subjects a 2 by 2 table defining all possible combinations of the hypothesis (true or false) with the evidence (supporting or not supporting the hypothesis). Such a table, among other things, allows the distinction between the two conditionals to be seen clearly. Each of these studies shows that with the help of this table, people do better at probabilistic inference word problems or at interpreting conditional probability statements.

A related possibility is that people might think that the two conditionals generally have the same numerical value. Pollatsek et al (1987) found that the majority of subjects believe this to hold for some problems.

There is some evidence against the Confusion Hypothesis. Pollatsek et al (1987) asked people for numerical estimates of both p(A/B) and p(B/A), for example, the probability that a person who has a fever is sick, and the probability that a person who is sick has a fever. They defined ranges of reasonable answers for each probability, and found little evidence that people reversed the two conditionals, i.e., gave answers for one that were reasonable for the other and vice versa. The form of Pollatsek et al's questions forces subjects to distinguish between the two conditionals, however, while the presentation in probabilistic inference word problems may allow them to be more easily confused.

5.2.1. Production system model.

The production system in Table 5 represents a set of plausible strategies by which a subject might use the p(E/H) number as if it were an appropriate answer to the question, "What is p(H/E)?" Rule 0 simply says that if the subjects already know the answer when asked the question, they will use it. Rule 2 says that if base rate information is available, and nothing else, then it will be used as the answer. Note that Rule 2 embodies the subjects' understanding of the applicability of the statistical information. As we showed above, the data support such a rule. Rule 1 says that if no information is given in the problem, the subjects will ask themselves to make an estimate of the statistical likelihood of the hypothesis, based on their own prior knowledge. This query establishes one of the conditions for Rule 8, which estimates base rate. In the absence of any statistical or relative frequency knowledge, Rule 8 sets the base rate estimate to .50. The conditions for Rule 2 are now met, and so the answer is set to the baserate estimate, .50. Note that Rules 3 and 4 embody a failure to think that the evidence might be unreliable. No reliability information (correctly interpreted or misinterpreted) has been given, and the subject believes the evidence 100%. Table 6 shows the sequence of rule applications for each information condition, and the predicted answers.

****
Insert Tables 5 and 6 about here.
****

It is not necessary in this model to specify what the subject would do if given evidence and reliability, or all three pieces of information, because the Confusion Hypothesis means that the subjects never recognize they are in one of these conditions. Specifically, whenever the subjects have evidence E and are given reliability p(E/H), they misinterpret the reliability information as p(H/E), and give .80 as the probability of the hypothesis. This is done by reading p(E/H) directly as p(H/E) and using Rule 0. (An alternative would be to add a rule

If
Query: p(A/B)

p(B/A)
Then
Set p(A/B) = p(B/A)

which represents the belief that the numerical value of the probability of A conditioned on B is the same as the probability of B conditioned on A. This rule would fire first, triggering Rule 0.)

The rules in this production system are sufficient to produce not only the answer .80 in the **b&e&r** condition, but also the most common answer in every condition. This may be seen by comparing Table 6 with Table 2.

5.3. Hypothesis 3: Integration of base rate and case information, with relative neglect of base rate.

The final explanation for the frequent use of the reliability, the Integration Hypothesis (Bar-Hillel, 1980; Tversky and Kahneman, 1982), assumes that the subjects do indeed understand the problem in terms of the conflicting pertinence of statistical information (the base rate) and case information (the evidence). They integrate or combine the two kinds of information in such a way that their final answer is between the two numbers, baserate (.15) and 1.0, and it is in fact closer to 1.0 than Bayes' Theorem would prescribe. Fischhoff and Bar-Hillel (1984) and Hamm (1987a) provide evidence supporting the prediction that the answer is between the base rate and 1.0. The bias toward 1.0 constitutes the "neglect" of base rate information. There is evidence supporting the notion that people neglect base rate information because they do not think it relevant. Bar-Hillel (1980) has shown that if the base rate information is made to seem more relevant, the answers shift toward it, as if it is given more weight in an integration process.

For perspective, let us look at how Bayes' Theorem would prescribe that the rate and evidence be combined. The interpolation between the numbers that represent full belief in the base rate and in the evidence depends on the reliability of the evidence. Thus, for the Cab problem, if the prior probability that a Blue Cab was involved is .15 and there is evidence that the unlikely event has happened, the lowest curve in Figure 1 shows how the correct degree of belief that the guilty cab was Blue is a function of the reliability. At the extreme values of reliability, p(H/E) = 1 if reliability p(E/H) = 1, and p(H/E) = base rate, if p(E/H) = .50. For the p(H) = .15 curve, the function is relatively flat between these extremes until reliability is fairly high. The other curves are for prior probabilities of .30, .50, .70, and .85.

****
Insert Figure 1 about here.
****

In contrast with the normative curve for p(H) = .15, people seem to act as if the dependency of p(H/E) on reliability is a function bowed in the other direction. Perhaps they are applying heuristics that are more appropriate for situations where the prior probability is greater than .50.

Although a biased interpolation process could account for p(H/E)'s that are higher than they ought to be, it is difficult for the Integration Hypothesis to account for the huge preference for the number .80. We need to assume that after interpolating between the base rate and 1.0, people select nearby available numbers (see Hamm, 1987a, 1987b). The process can now be represented as in Figure 2. This shows two mappings from reliability onto p(H/E) for a hypothesis with prior probability of .15: a normative mapping and the proposed subject mapping. The mapping on the left represents the correct interpolations between .15 and 1.0, for various reliabilities. It shows selected x-y pairs defined by the lowest curve in Figure 1. For example, if the reliability (on X axis in Figure 1) equals .80, then p(H/E) (on Y axis) equals .41.

****
Insert Figure 2 about here.
****

The mapping on the right represents the Integration Hypothesis view of how the typical subject

interpolates between base rate and 1.0. This differs in two ways from the normative interpolation. First, the subjects' mapping is more "straight across" than it ought to be. This produces the bias, and corresponds to a curve for p(H) of .15 that would be bowed the wrong way if plotted in Figure 1. Second, the mapping converges on available numbers -- 1.0, .80, .50, and .15 (see Table 2). Thus, it maps from ranges of reliabilities to a few corresponding points on the p(H/E) scale.

5.3.1. Production system model.

Table 7 shows a production system that represents the hypothesis that people integrate the statistical and case information, by using a biased discontinuous interpolation strategy to select a number between the .15 and 1.0. Rules 5, 6, and 9 represent the subjects' correct understanding that the problem requires the integration of statistical and case information. Rule 5 says that when the subject has information about evidence and its reliability, but lacks base rate information, he or she estimates it. Rule 6 calls for the interpolation between the statistical information, represented by the base rate, and the case information, represented by the 1.0 of complete confidence in the present evidence. Rules 9.1 to 9.4 do the interpolation, which is simply a mapping from four regions of the range of possible reliabilities onto the four available numbers. This production system accounts for the most common answers in the data, as shown in Table 8 (compare with Table 2).

****
Insert Tables 7 and 8 about here.
****

Note that the model is awkward and has has limited generality, because the mapping incorporated in Rule 9 is for a particular base rate, p(H) = .15. A different mapping would be required for problems with different base rates. However, these rules are adequate for the **e&r** condition, in which the base rate is .50.

6. Discussion.

Production system models of the three competing hypotheses were constructed in order to define subjects' heuristic strategies explicitly and to specify the conditions in which they are used. It was possible to eliminate one hypothesis about the source of subjects' use of reliability as the answer on probabilistic inference word problems, the idea that they universally ignore the base rate information on principle. The other two hypotheses, however, as represented by their production system models, are each consistent with the data. That is, each predicts the most common subject response in every one of eight information conditions in the study. This might appear unsurprising, because the models were written to do so. But it could have been very difficult to write a plausible model that was consistent with the data for one of these theories, just as it was for the Hypothesis of Principled Ignoring of Base Rate.

The natural direction for future work is to attempt to select between the hypotheses. However, consideration of individual differences in strategy and of task variation make an additional approach attractive, which would use multiple models of heuristic strategies, whose deployment would depend on individual and task factors.

6.0.1. Hypothesis comparison.

Selection between the Confusion and Integration hypotheses as explanations for subject performance on the Blue/Green Cab problem will depend on at least three factors. The first is their ability to predict answers on new, related problems. Application of these production systems to a different probabilistic inference word problem, about twins, has shown the Integration Hypothesis to be more successful (Hamm, 1987b). The second factor is the accuracy of the hypotheses' predictions of other evidence concerning the psychological strategies subjects use, e.g., verbalizations of judgment, memory search, or interpolation processes. Third, our judgments concerning the plausibility of the presumed psychological mechanisms play an important role in theory selection. In this regard, the way the interpolation process is represented in the Integration Hypothesis model, as a set of rules that perform stepwise mappings from p(E/H) to p(H/E) given a particular prior probability, is awkward and less plausible than the processes involved in the Confusion Hypothesis model. While this is consistent with the use of a

limited set of available numbers, observed in this study, it seems to require the prior existence of an unnecessarily large number of rules, particularly if we were to write a production system capable of handling many prior probabilities rather than just p(H) = .15.

6.0.2. Individual differences in strategies.

In testing hypotheses by comparing the predictions of their production system models with the data, we have considered only the most popular responses. This approach could reject a hypothesized strategy even though a minority of subjects might use it (see Chapter 10 of Hammond, McClelland, and Mumpower, 1980). For example, some subjects may indeed ignore the base rate on principle. Even though very few subjects gave answers consistent with Hypothesis 1 (Tables 2 and 4), it may be the perfect explanation for their answers. Further, if the problem were worded slightly differently, the plurality of subjects might act this way.

Although the Confusion and Integration hypotheses predict the same answers for all 8 conditions of this questionnaire study, they make distinct predictions for other possible observations, as noted above. Recognition of the possibility of individual differences suggests that future work should not blindly attempt to eliminate one or the other of these hypotheses. Rather, individual differences should be expected: some subjects may confuse p(E/H) with p(H/E), while others integrate the statistical and case information. A full account of subjects' answering of probabilistic inference word problems may require the description of a number of strategies and a counting of how many subjects use each strategy.

6.0.3. Task features influence subjects' strategies.

An additional aspect of this full account of answers to the Blue/Green Cab problem is a description of the task and how its possible variants might influence the strategies the subjects adopt (see Hammond, Hamm, Grassia, and Pearson, 1987). Let us assume that everyone integrates statistical and case information except when they misinterpret reliability information. Their tendency to confuse reliability p(E/H) with the probability of the hypothesis given the evidence p(H/E) may depend on particular aspects of the task. For example, Einhorn (in press; see also Einhorn and Hogarth, 1981) shows that the temporal relation between the terms in the conditionals may moderate the tendency to confuse them. In Eddy's (1982) example of a positive test result in a cancer screening, the cancer, if it is the cause of the positive test result, precedes the test in time; yet the test result is known first. Here the temporal cues to the interpretation of the meaning of the conditional probabilities are confusing, because both p(E/H) [p(positive test/cancer)] and p(H/E) [p(cancer/positive test)] can be seen as "the probability of a later event, given an earlier event." If one depends on temporal cues for help in interpreting conditional probabilities that one reads or hears, then word problems in which the temporal cues are ambiguous will be difficult.

A second factor that may contribute to confusion between p(H/E) and p(E/H) is that each of them can be properly called "accuracy" (Bar-Hillel, personal communication). As an illustration, if one's job is to evaluate AIDS tests, one would ask their success rates at detecting infected blood, p(E/H); but if one's job is to advise a worker who has tested positive on a screening, one would ask how accurate that test result is, p(H/E). Since both these conditional probabilities are commonly called "accuracy", it may be natural that people should confuse them in a word problem.

If, as is likely, the temporal cues for interpreting conditional probability statements, and the relative frequency with which the term "accuracy" is applied to the two conditional probabilities, vary according to the word problem content, then subjects' tendency to apply the strategies described by the Integration and Confusion Hypotheses will vary. To be prepared to handle such a task dependency, it would be better to develop each of these models, rather than to seek to reject one and declare the other the winner.

7. Bibliography.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. Acta Psychologica, 44, 211-233.

Bradshaw, Gary. (1987). Functional rules: Toward a synthesis of artificial intelligence and judgment/decision making. MORC Presentation, Institute of Cognitive Science, University of Colorado, Boulder.

Brownston, Lee, Farrell, Robert, Kant, Elaine, and Martin, Nancy. (1985). Programming expert systems in OPS5: An introduction to rule-based programming. Reading, Mass.: Addison-Wesley Publishing Company, Inc.

Christensen-Szalanski, J.J.J., and Beach, L.R. (1982). Experience and the base rate fallacy. Organizational Behavior and Human Performance, 29, 270-278.

Cohen, L.J. (1981). Can human irrationality be experimentally demonstrated? [with peer commentary] The Behavioral and Brain Sciences, 4, 317-370.

Dawes, R.M. (1986). Representative thinking in clinical judgment. Clinical Psychology Review, 6, 425-441.

Eddy, David M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, and A. Tversky, (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, pp 249-267.

Einhorn, Hillel J. (in press). Diagnosis and causality in clinical and statistical prediction. In D. C. Turk and P. Salovey (Eds.), Reasoning, Inference, and Judgment in Clinical Psychology. New York: The Free Press.

Einhorn, Hillel J., and Hogarth, Robin M. (1981). Uncertainty and causality in practical inference. Center for Decision Research, Graduate School of Business, University of Chicago.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1979). Subjective sensitivity analysis. Organizational Behavior and Human Performance, 23, 339-359.

Fischhoff, B., and Bar-Hillel, M. (1984). Focusing techniques: A shortcut to improving probability judgments? Organizational Behavior and Human Performance, 34, 175-194.

Hamm, Robert M. (1987a). Diagnostic inference: People's use of information in incomplete Bayesian word problems. Institute of Cognitive Science Publication #87-11, University of Colorado, Boulder.

Hamm, Robert M. (1987b). A model of answer choice on probabilistic inference word problems. Society of Mathematical Psychology meetings, Berkeley, CA, August, 1987.

Hammond, K.R., Hamm, R.M., Grassia, J., and Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. IEEE Transactions on Systems, Man, and Cybernetics, SMC-17, 753-770.

Hammond, K.R., McClelland, G.H., and Mumpower, J. (1980). Human Judgment and Decision Making: Theories, methods, and procedures. New York: Praeger.

Johnson, E.J., and Payne, J. (1985). Effort and accuracy in choice. Management Science, 30, 1213-1231.

Kahneman, D., and Tversky, A. (1972b). On prediction and judgment. Oregon Research Institute

Research Monograph, 12(4). Cited in Fischhoff and Bar-Hillel, 1984.

Lichtenstein, S., and MacGregor, D. (1984). Structuring as an aid to performance in base-rate problems. Report #84-16, Decision Research, Eugene, Oregon.

Lyon, D., and Slovic, P. (1976). Dominance of accuracy information and neglect of base rates in probabiity estimation. Acta Psychologica, 40, 287-298.

Niiniluoto, I. (1981). L. J. Cohen versus Bayesianism. The Behavioral and Brain Sciences, 4, 349.

Pollatsek, Alexander, Well, Arnold D., Konold, Clifford, Hardiman, Pamela, and Cobb, George. (1987). Understanding conditional probabilities. Organizational Behavior and Human Decision Processes, 40, 255-269.

Simon, H.A. (1976). Discussion: Cognition and social behavior. In J.S. Carroll and J.W. Payne (Eds.), Cognition and Social Behavior. Hillsdale, N.J.: Erlbaum, pp 253-267.

Tversky, A., and Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, and A. Tversky, (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, pp 153-160.

Widiger, T.A., Hurt, S.W., Frances, A., Clarkin, J.F., and Gilmore, M. (1984). Diagnostic efficiency and DSM-III. Archives of General Psychiatry, 41, 1005-1012.

8. Tables

Table 1. The Cab Problem used in this study, with the separate parts identified.

Introduction. The next word problem is about two taxi cab companies. A cab from one of the companies was involved in a hit and run accident at night. It is hard to know which company it was from. You will be asked to estimate how likely it is that the cab involved in the accident belonged to each of the two cab companies.

In this city there are only two cab companies, the Blue Cab Company and the Green Cab Company.

With what you know now, what is the probability that the cab involved in the hit and run accident was from the Blue Cab Company? _____

Evidence. There was only one witness to the hit and run accident. The witness identified the cab as blue.

With what you know now, what is the probability that it was a Blue Cab? _____

Base rate. The Green Cab Company is larger, with 85% of the cabs in the city.

With what you know now, what do you think is the probability that a cab from the Blue Cab Company was the one involved in the accident? _____

Reliability. The police were concerned about the accuracy of the witness who saw the accident. They tested the witness's reliability under the same circumstances that existed on the night of the accident and concluded that the witness could correctly identify cabs of each one of the two colors 80% of the time and misidentified them 20% of the time.

With what you know now, what is the probability that the cab was a Blue Cab? _____

Table 2.
Number of subjects responding using each available number, for each condition (combination of presented information). Blue/Green Cab problem.

	Condition							
Answer:	No Info.	Basert only	Evid. only	Reliab only	Basert & Rel.	Basert & Evid	Evid. & Rel	Basert, Evid & Rel
.15 Baserate	0	40	0	0	30	6	0	12
.5 Equal chance	252	2	10	65	5	10	3	10
.80 Reliability	0	0	6*	8	2	4*	64	101
1.0 (Full faith in evidence)	0	0	33	0	0	19	1	4
Other Numbers	4	41	36	14	48	44	16	120
Total	256	83	85	87	85	83	84	247

* In these cells, the row information has not yet been presented.

Table 3:
Production System for Hypothesis of Principled Ignoring of Base Rate. Blue/Green Cab Problem.

Rule No.	Conditions: If These Are True:		Actions: Then Do This:
1	Query: p(H)	--->	Answer = .50
3	Query: p(H/E) Evidence	--->	Answer = 1.0
5	Query: p(H/E) Evidence Reliability	--->	Set Query: Calculate Bayes' Theorem
12	Query: Calculate Bayes' Theorem Evidence Reliability	--->	p(H/E) = reliability (see text)

Table 4.
Sequence of Production Applications for Principled Ignoring of Base Rate Hypothesis Production System.

Information Condition	Predicted Answer	Rule Sequence
None	.50	1
b	.50	1
e	1	3
r	.50	1
b&e	1	3
b&r	.50	1
e&r	.80	5-12
b&e&r	.80	5-12

b = Baserate; e = Evidence; r = Reliability

Table 5.
Production System for Confusion Hypothesis.
Blue/Green Cab Problem.

Rule No.	Conditions: If These Are True:		Actions: Then Do This:
0	Query: p(H/E) Evidence p(H/E)	--->	Answer = p(H/E)
1	Query: p(H)	--->	Set Query: Estimate Baserate
2	Query: p(H) Baserate	--->	Answer = Baserate
3	Query: p(H/E) Evidence	--->	Answer = 1.0
4	Query: p(H/E) Baserate Evidence	--->	Answer = 1.0
8	Query: Estimate Baserate No Statistical Info	--->	Baserate = .5

Table 6.
Sequence Of Production Applications For Confusion Hypothesis Production System.

Information Condition	Prediction & Most Common Answer	Rule Sequence
None	.5	1-8-2
B	.15 (= B)	2
E	1	3
R	.5	1-8-2
B&E	1	4
B&R	.15 (= B)	2
E&R	.80 (= R)	0
B&E&R	.80 (= R)	0

B = Baserate; E = Evidence; R = Reliability

Table 7.
Production System for Integration Hypothesis.
Blue/Green Cab Problem.

Rule No.	Conditions* If These Are True:		Actions Then Do This:
0	Evidence p(H/E)	--->	Answer = p(H/E)
1	Null	--->	Set Query: Estimate Baserate
2	Baserate	--->	Answer = Baserate
3	Evidence	--->	Answer = 1.0
4	Baserate Evidence	--->	Answer = 1.0
5	Evidence Reliability	--->	Set Query: Estimate Baserate
6	Baserate Evidence Reliability	--->	Set Query: Interpolate p(H/E)
8	Query: Estimate Baserate No Baserate Info	--->	Adopt Baserate = .5
9.1	Query: Interpolate p(H/E) B, E, And R R Is Very High ($R \geq .95$)	--->	p(H/E) = 1.0
9.2	Query: Interpolate p(H/E) B, E, And R R Is High ($.75 \leq R < .95$)	--->	p(H/E) = .80
9.3	Query: Interpolate p(H/E) B, E, And R R Is Low ($.55 \leq R < .75$)	--->	p(H/E) = .50
9.4	Query: Interpolate p(H/E) B, E, And R R Is Very Low ($R < .55$)	--->	p(H/E) = Baserate

*In addition to the conditions listed, each production has the condition "Query for p(H)" or, if

evidence is available, "Query for p(H/E)".

Table 8. Sequence of Production Applications for Integration Hypothesis Production System.

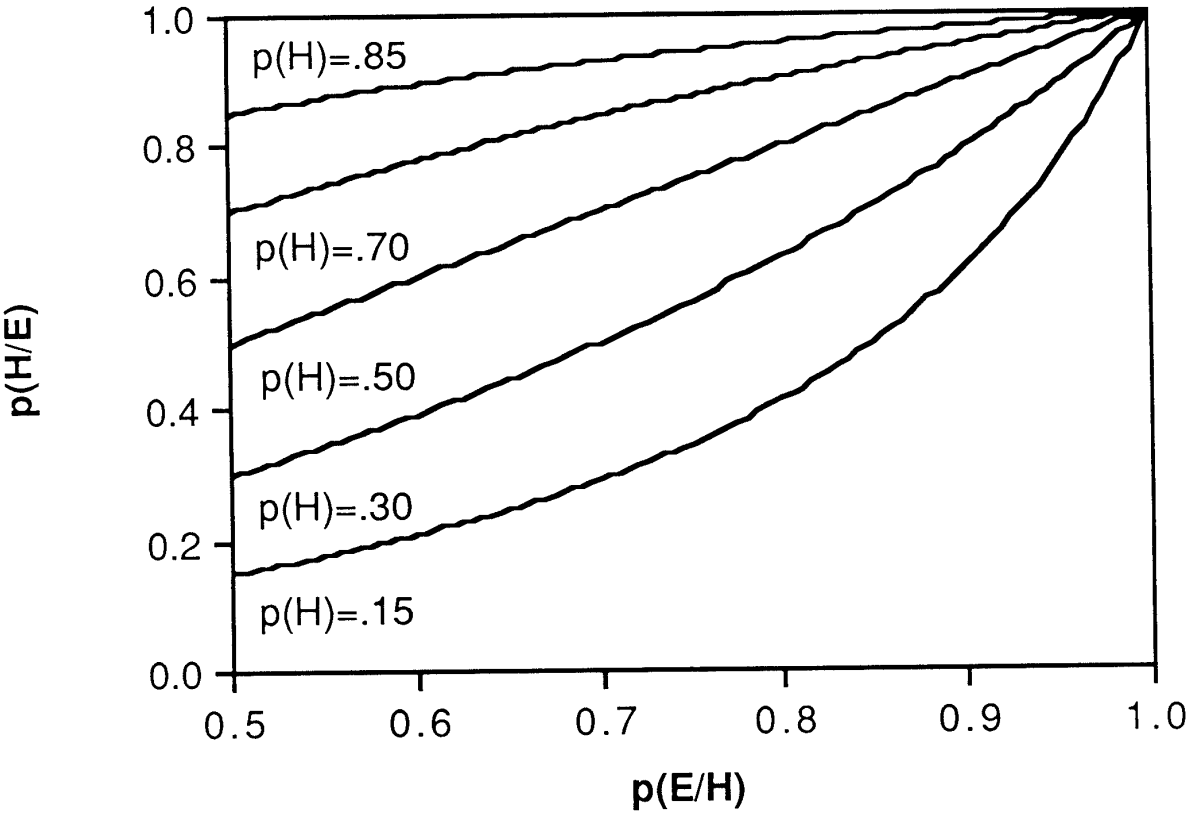
Information Condition	Pred. Answer	Rule Sequence
0	.5	1-8-2
b	.15	2
e	1	3
r	.5	1-8-2
b&e	1	4
b&r	.15	2
e&r	.80	5-8-6-9.2
b&e&r	.80	6-9.2

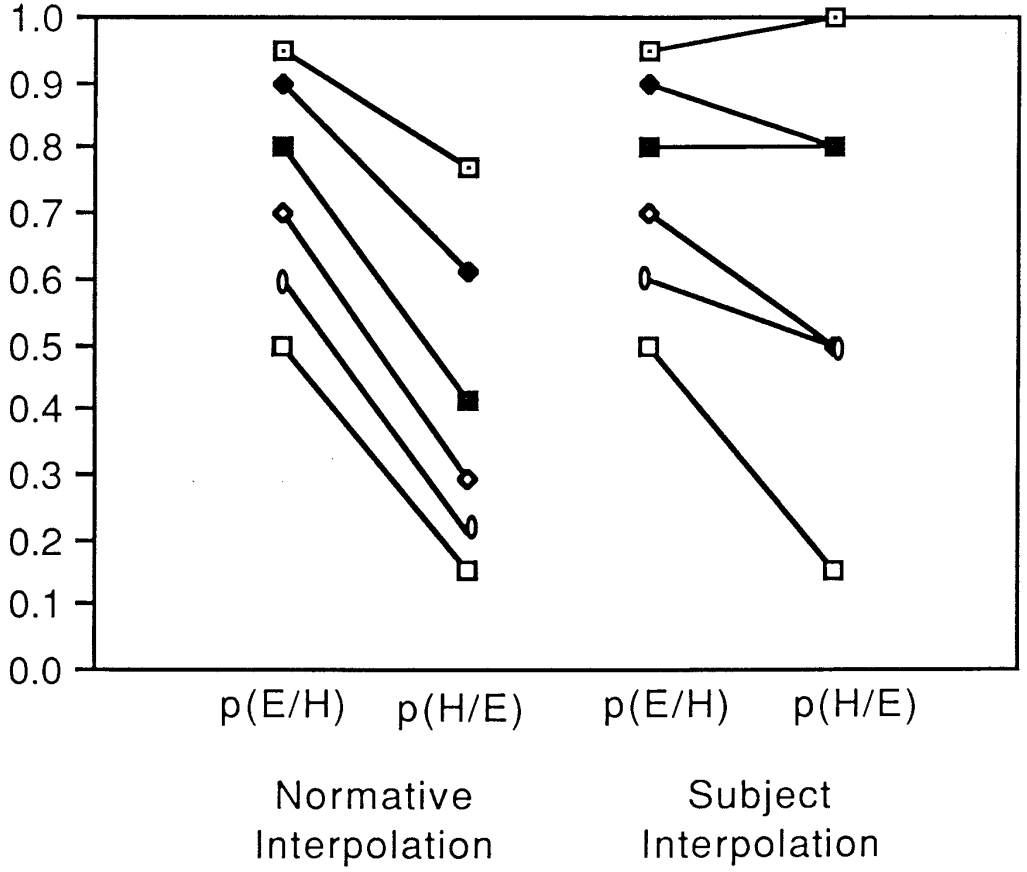
b = baserate; e = evidence; r = reliability

9. Figure Captions.

Figure 1. Degree of belief in hypothesis give evidence, p(H/E), as a function of reliability, p(E/H), for each of several prior probabilities.

Figure 2. Integration of statistical and case information to produce p(H/E) by interpolation between base rate and 1.0, as a function of p(E/H), for prior p(H) = .15.


1.0
0.8
0.6
0.4
0.2
0.0
p(H/E)
p(H)=.85
p(H)=.70
p(H)=.50
p(H)=.30
p(H)=.15
0.5
0.6
0.7
0.8
0.9
1.0
p(E/H)


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
p(E/H)
p(H/E)
p(E/H)
p(H/E)
Normative Interpolation
Subject Interpolation