Interpretation of Conditional Probabilities in Probabilistic Inference Word Problems.

Robert M. Hamm and Michelle A. Miller

Institute of Cognitive Science
Box 345
University of Colorado,
Boulder, Colorado, 80309-0345.
303/492-2936

Institute of Cognitive Science
Publication Number 88-15

December 1988

Running head: INTERPRETATION OF CONDITIONAL PROBABILITIES

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): 88-15	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION: Institute of Cognitive Science; 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State and ZIP Code): University of Colorado, Box 345, Boulder CO 80309-0345	7b. ADDRESS (City, State and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: Army Research Institute, BR; 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: MDA-903-86-K-0265
8c. ADDRESS (City, State and ZIP Code): 5001 Eisenhower Avenue PERI-BR, Alexandria VA 22333-5600	10. SOURCE OF FUNDING NOS. — PROGRAM ELEMENT NO.; PROJECT NO.; TASK NO.; WORK UNIT NO.
11. TITLE (Include Security Classification): Interpretation of Conditional Probabilities in Probabilistic Inference Word Problems.	
12. PERSONAL AUTHOR(S): Robert M. Hamm and Michelle A. Miller	
13a. TYPE OF REPORT: Technical; 13b. TIME COVERED: FROM _____ TO _____	14. DATE OF REPORT (Yr., Mo., Day): December 30, 1988; 15. PAGE COUNT
16. SUPPLEMENTARY NOTATION: Contracting Officer's Representative was Michael Drillings.	

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB. GR.	Protocol Analysis, Conditional Probability, Probabilistic Inference, Bayes' Theorem, Expert/Novice

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

People's strategies on probabilistic inference word problems were investigated in an attempt to determine which of 3 theories explains their neglect of base rate information when estimating the probability of a hypothesis. These word problems present a base rate or prior probability p(h), some evidence e which typically conflicts with the prior expectation, and information on the reliability of the evidence which is stated as p(e/h), the conditional probability of the evidence being seen if the hypothesis is true.

The three theories are (a) subjects believe that the base rate is irrelevant, (b) they integrate base rate and evidence in a manner which happens to underweight the base rate, or (c) they misinterpret the reliability information p(e/h) as if it were p(h/e). Data using four distinct methods support the theory that subjects confuse the conditional probabilities.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION
22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE NUMBER (Include Area Code); 22c. OFFICE SYMBOL

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

SECURITY CLASSIFICATION OF THIS PAGE

Table of Contents

Interpretation of Conditional Probabilities in Probabilistic Inference Word Problems.

Abstract.

People's strategies on probabilistic inference word problems were investigated in an attempt to determine which of three theories explains their neglect of base rate information when estimating the probability of a hypothesis. These word problems present a base rate or prior probability *p(h)*, some evidence *e* which typically conflicts with the prior expectation, and information on the reliability of the evidence which is stated as *p(e/h)*, the conditional probability of the evidence being seen if the hypothesis is true.

The three theories are (a) subjects believe that the base rate is irrelevant, (b) they integrate base rate and evidence in a manner which happens to underweight the base rate, or (c) they misinterpret the reliability information *p(e/h)* as if it were *p(h/e)*. Data using four distinct methods support the theory that subjects confuse the conditional probabilities. Substitution of *p(h/e)* for *p(e/h)* in the stimulus problems caused no changes in the distribution of subjects' answers. When subjects were offered a list of types of information to use in solving the problem, including *p(e/h)* and *p(h/e)*, their preferences were consistent with a strategy of requesting both conditional probabilities together because they do not know which one they need. Analysis of subjects' use of conditional probability concepts when thinking aloud about the reliability information they had been given showed that their use of *p(e/h)* or *p(h/e)* concepts was almost independent of which one they were exposed to. When asked unexpectedly to recall the key information from a problem, subjects frequently recalled the base rate, and often reported the opposite conditional probability from the one originally presented.

If the neglect of base rate information is due to misinterpretation of the conditional probabilities expressing the reliability of the evidence, then programs designed to improve performance should focus as much on the reliabilities as on the base rates.

1. Introduction.

This paper investigates an aspect of people's reasoning about situations in which they must integrate (a) information about *what usually happens* and (b) imperfectly reliable evidence about *what is happening in the present case*, to estimate the degree to which they can believe their hypotheses about what is happening in the present case. For example, intelligence analysts confronted with unreliable evidence that suggests a very unusual pattern of enemy activity would need to make a formal or informal estimate of how likely it is that this unusual activity is indeed happening, as a basis for deciding whether to issue an alert or just investigate further. The aspect of interest is people's interpretation of information about the general relation between evidence and hypotheses in such situations. This kind of information can be presented as a conditional probability: the probability that the presently observed evidence would be seen if a hypothesis were true. However, people don't know exactly what to do with this kind of information, and they frequently act as if it means the same thing as its opposite, the probability that a hypothesis would be true if particular evidence were seen.

Distinguishing these two kinds of conditional probability is crucial for the successful application of probability theory to the problem of inference. When observable evidence is only probabilistically associated with the states of the world that we care about, then it is appropriate to deal with degrees of belief, rather than with certainties. In such uncertain situations, inference is the adjustment of degrees of belief in hypotheses, given new evidence. If we can use probabilities to measure our beliefs in hypotheses, as well as the associations between the hypotheses and the evidence, then Bayes' Theorem offers a method for adjusting the degree of belief when we are given new evidence (Edwards, Lindman, and Savage, 1963). While there is controversy about the appropriateness of Bayes' Theorem versus alternative formulations of probabilistic inference (L. J. Cohen, 1981; M. Cohen, et al, 1985; Schum, 1987), still in many situations Bayes' Theorem is a reasonable standard. As such, it is of interest to compare people's reasoning about realistic problems to this standard. How do people's probabilistic inferences differ from Bayes' Theorem inferences? Even if quantitatively inaccurate, do people use information in the same qualitative manner as Bayes' Theorem?

Two bodies of research have studied people's unaided probabilistic inference. The first is the book-bag and poker chip paradigm (reviewed by Edwards, 1968, and Slovic and Lichtenstein, 1971), which investigated peoples' combination of information about a prior probability (summary of beliefs before new observations) plus multiple competing unreliable observations. This approach compared people's performance to the odds-likelihood form of Bayes' Theorem. The second approach is the diagnostic word problem paradigm (reviewed by Tversky and Kahneman, 1982), which investigated people's combination of a prior probability with a single unreliable observation that is inconsistent with the prior expectation. This approach compared people's performance to the simple form of Bayes' Theorem. The present work addresses the latter problem. While some have concluded that the book-bag and poker-chip approach compares people with too difficult a standard (von Winterfeldt and Edwards, 1986), the diagnostic word problems of the second approach are similar to situations in which most people occasionally find themselves, and with which some experts must deal on a regular basis, with important consequences (e.g., medical doctors; see Eddy, 1982; and intelligence analysts; see Cohen et al, 1985, and Schum, 1987). As such, it is worth while to study how probabilistic inference is done by people who have different degrees of experience with the formal standards of probabilistic inference.

We next review standards for probabilistic inference, then compare people's behavior on probabilistic inference word problems to these standards and present theories that explain this behavior.

1.1. Standards for probabilistic inference.

When people's numerical responses to probabilistic inference word problems are compared to the precise standard defined by the mathematics of Bayes' Theorem, they always come up short. It is more instructive to compare their behavior to a qualitative standard, that is, to see whether their responses have the same qualitative relations to the key variables in probabilistic inference word problems as Bayes' Theorem's answers do.

1.1.1. The quantitative Bayes' Theorem standard.

Appendix I presents the probabilistic inference word problems used in this research. The issue in the first problem is which of two twin boys is the one who was seen knocking over a lamp. The babysitter thought it was Steve, but was not sure. There is sufficient information in the story to allow the application of Bayes' Theorem.

$$p(h/e) = \frac{p(h)\times p(e/h)}{p(h)\times p(e/h) + p(\sim h)\times p(e/\sim h)}$$

The probability that it really was Steve who broke the lamp, given the baby sitter thought it was Steve, p(S/"S"), is a complicated function (a ratio of sums of products) of the prior or base rate probability that Steve would break a lamp (an instance of being a trouble maker), p(S) or .20, the prior probability for the other twin Paul, p(P) or .80, the probability that the babysitter would think Steve was Steve, p("S"/S) or .60, and the probability she would think Paul was Steve, p("S"/P) or .40. Specifically,

$$p(Steve/"Steve") = \frac{p(Steve)\times p("Steve"/Steve)}{p(Steve)\times p("Steve"/Steve) + p(Paul)\times p("Steve"/Paul)}$$

1.1.2. The qualitative relations consistent with the Bayes' Theorem standard.

If people do not both know Bayes' Theorem and have computational tools, then exact application of the formula can not be expected (von Winterfeldt and Edwards, 1986). Nonetheless, we can inquire whether their behavior has qualitative features that are consistent with the prescriptions of Bayes' Theorem. These are, at minimum, that the impacts of each kind of relevant information be in the right direction. Specifically, for the two-hypothesis case,

1. If there is evidence, then it ought to make the subject consider the hypothesis it supports to be more likely.
2. If there are good *a priori* reasons to believe a hypothesis, then the higher this prior probability, the more likely the subject should consider that hypothesis to be; as a special case, if there is relative frequency information about what usually happens, then the higher this base rate, the more likely the subject should consider the hypothesis to be.
3. If in addition to evidence, there is information about the reliability of the evidence, such as information about how frequently the particular evidence would be seen if the hypothesis that the evidence points to were true, then the higher this conditional probability, the more likely the subject should consider the hypothesis to be.
4. If in addition to evidence, there is a second type of information about the reliability of the evidence, which is how frequently the particular evidence would be seen if the complementary hypothesis (the one that the evidence seems to contradict) were true, then the higher this conditional probability, the less likely the subject should consider the hypothesis to be.
5. If evidence is accompanied by both prior probability (or base rate) information and information about the relation between the hypothesis and the evidence, then all this information should be integrated.
6. It is possible to make reasonable assumptions about information that has not been specified.

These qualitative relations are also required by the alternatives to Bayes' Theorem, with the possible exception of Cohen's "inductive probabilities" (Cohen, 1977; see Schum, 1987).

1.2. People's behavior on probabilistic inference word problems.

Research on probabilistic inference has almost universally found that people's numerical answers differ from those produced by applying Bayes' Theorem. Even when subjects' median answer is very accurate, researchers have concluded that people are not actually calculating Bayes' Theorem, and are producing "fairly close to optimal answers for the 'wrong' reasons" (Ofir, 1988, p 361).

Researchers have emphasized three features of people's probabilistic inferences. Bar-Hillel (1980), Kahneman and Tversky (1972; Tversky and Kahneman, 1982), and others have argued that people neglect the base rate information. Doherty, Mynatt, Tweney, and Schiavo (1979) and Beyth-Marom and

Fischhoff (1983) have said that people underutilize or ignore the false alarm information, p(e/~h), the probability that the evidence that favors hypothesis h could have been seen if h were not true. Finally, Hamm (1987a, 1988a) has argued that people use numbers available in the word problems as their answers, and thus are not integrating the information about base rate and the reliability of evidence as they ought. Let us consider these findings in detail so we can know what must be explained by a theory of human probabilistic inference.

1.2.1. The neglect of base rate.

Kahneman and Tversky (1972), Lyon and Slovic (1976), Bar-Hillel (1980), and many others have shown that on word problems with a low base rate [for example, p(h) = .15], a moderate reliability [e.g., p(e/h) = .80], and the complementary false alarm rate [p(e/~h) = .20], subjects' answers are too high. Where the answer calculated for these particular values using Bayes' Theorem is .41, subjects' median and modal answer is .80 (Kahneman and Tversky, 1972; Bar-Hillel, 1980). This does not mean, however, that people completely ignore base rate information.

Several studies have shown that people's answers on these problems vary in response to base rate. When only the base rate was presented, without evidence, people used it as their answer (Lyon and Slovic, 1976; Hamm, 1987a; Ofir, 1988a). When the base rate was made to seem causally connected to the present case, people paid more attention to it (Bar-Hillel, 1980). Studies which varied the level of base rate in repeated presentations of the same problem showed that people respond to it (Fischhoff, Slovic and Lichtenstein, 1979; Birnbaum and Mellers, 1983; Ofir and Lynch, 1984). However, Fischhoff and Bar-Hillel (1984) cautioned that the experimental situation may have demanded such a response: what else is there for the subject to respond to, in repeated versions of the same problem, but the numbers that are varying in the problems? In reply, Ofir and Lynch (1984) and Ofir (1988) varied the base rate in between-subjects designs and found that the mean and median answers were usually responsive to base rate differences. Ofir (1988) found that base rate was attended over the range of possible hit rates p(e/h) and false alarm rates p(e/~h), except when hit rate was high and false alarm rate low.

Thus only when the hit rate and false alarm rate both support the evidence, by being high and low, respectively, is competing base rate information neglected, in violation of the second qualitative Bayesian principle[1]. The problems studied by Kahneman and Tversky (1972), Bar-Hillel (1980), and others have just these characteristics. Although these problems represent only a subset of the possible probabilistic inference word problem variants, it is an important subset. These are situations where the evidence is inconsistent with prior expectations. It remains to be explained why people attend the evidence in probabilistic inference word problems, while they may be governed by their expectations or prejudices in other situations. (Though this has been attributed to the salience or perceived relevance of the base rate, that explanation is nearly tautological.)

1.2.2. The neglect of false alarm information.

To determine whether people neglect information concerning the possibility that the alternative hypothesis is true and has been mis-identified, Ofir (1988) presented problems with a range of p(e/~h) values and found that in most cases subjects' answers vary in the appropriate direction in response to these variations. The exceptions are that false alarm information seems to be ignored either when the false alarm rate is very low, or when both the base rate and the hit rate are very high. These qualitative inconsistencies with Bayes' Theorem (Principle 4) seem to be due to simplifying strategies -- completely ignoring variables that have a small but important impact. As such, these deviations do not require special explanation. Ofir (1988) additionally noted, following Beyth-Marom and Fischhoff (1983), that while people know how to use the false alarm information if they have it, they do not think to look for it if it has not been given to them. This conflicts with the special sixth qualitative Bayesian property listed above: that people should make reasonable guesses at missing information.

1.2.3. The failure to integrate competing base rate and evidence information.

Hamm (1987a) found that a majority of subjects responded to these word problems using one of the numbers that are available in the problem. For example, when evidence, base rate, and the reliability of evidence are all available, each itself was used by some subjects (1.0 is the number associated with the

exclusive use of the evidence), and the hit rate p(e/h) was the most popular response. None of these is a correct answer, and their use suggests people's strategies differ qualitatively from the fifth Bayes' Theorem standard. That is, they do not integrate the two competing kinds of information, evidence and base rate.

Hamm (1988a) did a study to test whether the over-estimation of the probability of the hypothesis, in probabilistic inference word problems with low base rate and high hit rate, might occur because the numerical probabilities in the problem are unfamiliar and induce people to adopt simplifying strategies, such as selecting available probabilities, that they would not normally use. Verbal expressions of probability were substituted for the numerical expressions in the problems, and subjects responded verbally. Although subjects in the verbal probability condition selected the probabilities available in the word problem as their answer much less frequently than subjects in the numerical probability condition, there was still a substantial under-utilization of the base rate information.

Our review of the qualitative features of people's performance on probabilistic inference word problems has shown that (a) although in most situations their responses covary appropriately with changes in the key base rate and reliability information, people do not seem to be integrating the cues, particularly when the cues have competing implications; (b) people's numerical responses are sometimes quite far from the Bayes' Theorem answer even when they covary appropriately with all inputs (see figures in Ofir, 1988). This means that those who must submit their health or security to decision systems that rely on people's intuitive probabilistic inference have something to worry about. And those of us who venture to offer advice on these decision systems need to understand what people are doing on these tasks.

1.3. Theories of people's probabilistic inference behavior.

Three theories of how people do probabilistic inference word problems offer explanations of the frequent use of the reliability or hit rate, p(e/h), as the response. The theories hold that people (a) view the base rate p(h) as irrelevant, (b) integrate p(h) and p(e/h) by interpolating, but interpolate inaccurately, or (c) misinterpret the reliability of the evidence p(e/h) as the opposite conditional probability p(h/e). Precise models of each of these theories, stated as production systems, are in Hamm (1987b). A fourth theory, a generalization of the second, will be explained here, but will not explicitly be tested in the study.

1.3.1. Purposeful neglect of base rate.

Cohen (1981) considered the apparent neglect of the base rate information and argued that the everyday intuition of the common (i.e., not specially educated) person could not be judged to be "wrong", because ordinary reasoning "sets its own standards" (p 328). Therefore he supposed that people have good reasons to neglect the base rate, and he ventured to articulate these reasons for them. A key factor is whether the prior probability pertains to the specific case or is just derived from information about relative frequencies. In Cohen's view, relative frequency probabilities are not pertinent to the probability of a hypothesis being true for a unique case. In a unique case, "we have to suppose equal predispositions" (p 329). The answer should therefore be determined by the probability that the evidence is false.

The theory of how people make probabilistic inferences when they are purposefully neglecting the base rate can be generalized from Cohen's example of how the reasonable person responds on the Blue/Green Cab problem (Cohen, 1981, pp 328-329). The decision maker is a juror; the issue is the probability that the cab that was involved in a hit and run accident was a blue one. The only other alternative color is green. A witness said the cab was green. Cohen says that jurors should be concerned just with "the probability that the cab actually involved in the accident was blue, on the condition that the witness said it was green" [p(~h/e), in our generic notation]. And the pertinent information is just the fact that the court is told that the witness who called it green "can distinguish blue cabs from green ones in 80% of cases" [p(e/h) = p(~e/~h) = .80; therefore, p(~e/h) = 1 - p(e/h) = .20]. Thus, "if the jurors know that only 20% of the witness' statements about cab colours are false, they rightly estimate the probability at issue as 1/5, without any transgression of Bayes' law" (p 328).

Thus Cohen asserts that the common person's right answer to the question "what is p(~h/e)?" is "p(~e/h)", or equivalently, the answer to "what is p(h/e)?" is "p(e/h)". To arrive at this answer, the common

person first assumes that the relative frequency or base rate is not pertinent, and then takes the probability that a witness' statement about a cab color is false, which is available, to be the probability at issue, that is, the probability that the cab in question was blue, if the witness said it was green. This last step involves appropriating an available p(e/h) for the desired p(h/e). This is the "confusion" that defines the Confusion Theory. However, the Confusion Theory does not require that base rate be ignored, as this theory does. Further, Niiniluoto (1981), in support of the theory that people purposefully neglect base rate, offers an alternative route to the p(e/h) answer that does not involve a confusion of conditional probabilities. If the base rate information is assumed to be irrelevant, and one therefore adopts a prior probability of .50, then the application of Bayes' Theorem produces the answer p(h/e) = p(e/h) (see Hamm, 1987b). Thus we have two distinct versions of a descriptive theory that says people ignore the base rate in these conditions because they reasonably judge it to be irrelevant: one where they also confuse the given p(e/h) for the desired p(h/e), and one where they apply Bayes' theorem accurately and discover that the needed p(h/e) equals the given p(e/h).

We note that the more recent studies reviewed above, showing the utilization of base rate in most situations, do not support the basic assumption of this theory. However, Cohen (1981, p 329) qualifies his assertion that base rate should be ignored, stating that in the absence of any specific evidence, it might be reasonable to attend to it. Therefore we keep it as a viable hypothesis for this study.

We also note that the theory that statistical base rates are not relevant to unique cases is applicable only to word problems where there is no specific connection to the particular case. Therefore it would not apply to our Twins problem (see Appendix I), where the base rate information is about the particular individuals, though it would apply to our Doctor and Insurance problems.

1.3.2. Interpolation.

The second theory is that subjects do indeed appreciate the pertinence of both the base rate and the unreliable evidence. They try to integrate these, but their integration is inaccurate. The method they use may be weighted averaging, anchoring and adjustment, a generic interpolation process, or even a rough application of Bayes' Theorem. The integration typically gives less weight to the base rate than is appropriate (Bar-Hillel, 1980; Tversky and Kahneman, 1982). Finally, people may round off the result of their integration process, to a nearby probability landmark (from the set including .90, .80, etc.), or to the nearest number that is available in the problem (Hamm, 1987c).

1.3.3. Confusion.

The third theory holds that the conditional probability expressing the reliability of the evidence [p(e/h), the probability of observing a particular piece of evidence given a particular hypothesis is true], is used as the response because people confuse it with the conditional probability [p(h/e), the probability that the hypothesis is true given the evidence has been observed]. The problem asks for the latter, but people think the former is an appropriate answer (Eddy, 1982; Dawes, 1986).

As we demonstrated above, Cohen (1981) attributed this confusion to the common person, in justifying the purposeful neglect of base rate, and so the confusion of conditional probabilities is a part of the Cohen (but not the Niiniluoto) version of that theory. However, it is possible to adopt p(e/h) as the answer to the probabilistic inference problem even if one considers base rate information to be pertinent. If one either believes that p(e/h) is equal to p(h/e), or does not recognize that they are two different concepts and identifies p(e/h) as the needed p(h/e), then one does not need to consider base rate at all, because apparently the needed answer is immediately available. Such a belief would also account for people not attempting to integrate the base rate with the unreliable evidence.

1.3.4. Cue utilization contingent on cue conflict.

Ofir (1988) proposed that people solving probabilistic inference word problems attend to different numbers of cues depending on whether the cues have converging or conflicting implications. If the small number of cues the subject first attends to is internally inconsistent, he or she will pay attention to another cue, searching for a strong conclusion. We do not consider this theory in the present study, because: (a) it does not specify the order in which the cues are attended; (b) it does not specify the rules governing the

production of the answer, once all cues have been attended; and (c) it does not specify the level of each cue that is considered as support or conflict: the one specific prediction, which is that p(e/~h) is considered to support the evidence if below .50 and to contradict the evidence if above .50, is not confirmed by discontinuities at .50 that would have been expected in the data curves reported by Ofir (1988).

Although the three theories describe different processes that people use on probabilistic inference word problems, they predict the same answers. In an effort to broaden the span of the theories to increase the chances of falsifying them, Hamm (1987b) looked at people's answers when they had been presented with each possible subset of the key information in these problems. However, it was possible, for each of the three theories, to produce a model that predicted the subjects' most common response for every possible combination of information.

We need, therefore, to observe behavior other than subjects' answers in order to discriminate among the theories. In this study, then, in addition to the usual analysis of answers, we make use of process tracing, protocol analysis, and memory techniques, in conjunction with experimental manipulation of the stimuli. Specifically, the process tracing technique (Payne, 1976) gives people the opportunity to choose which information to receive. This can reveal whether they value base rate information, and whether they discriminate between p(e/h) and p(h/e) information. The protocol analysis technique (Ericsson and Simon, 1984) requires people to think aloud while answering the questions and while choosing which information to receive. Analysis of their verbalizations can reveal the strategies they use, the information that they pay attention to, and the distinctions they make between the conditional probabilities. The memory technique requires subjects to recall a problem and their response, after a delay. The concepts they recall can reveal how they represented the information given in the problem (cf Kozminsky, Kintsch, and Bourne, 1981; Dellarosa and Bourne, 1984; McClelland, Stewart, Judd, and Bourne, 1987). To investigate the Confusion hypothesis, the reliability information in the word problems was varied systematically between p(e/h) and p(h/e).

2. Method.

Subjects were asked to solve three word problems, to think aloud while working on the problems, and to recall the first problem after completion of the third.

2.1. Subjects.

Subjects were 16 undergraduates (7 males) who responded to advertisements posted in the psychology department and were paid with either money or subject pool course credit, 14 mathematics or engineering graduate students (11 males) recruited by phoning lists of students provided by their departments, and 3 insurance professionals. The graduate students were screened to assure that they had had at least two courses in probability, in order to find people who had learned Bayes' Theorem. Because the issue of probability training had been raised, the graduate students were asked not to review probability before they came in. The researcher did not mention Bayes' Theorem in the recruiting conversation.

The primary purpose for studying these different groups was to compare people who were and were not trained in the mechanics of probabilistic inference. The insurance professionals were included to explore differences due to experience with the Insurance problem, and therefore they were not required to complete all three problems unless they volunteered. Due to problems with tape recorders, it was not possible to produce verbal protocols for every subject who completed a questionnaire.

2.2. Word problems and problem types.

Three different problems were used, the Doctor problem and Twins problem (used by Hamm, 1987a), and the Insurance problem (see Appendix I). The "facts" in each problem are plausible fictions. Each problem was prepared in two versions: with a *fixed* order and a *subject-selected* order of presentation of information.

In the fixed order version, the problem was divided into four paragraphs, which presented (1) an introduction to the problem, which identified the two mutually exclusive hypotheses, (2) base rate information, (3) evidence, and (4) conditional probability information. The subject was asked to estimate the probability of the hypothesis after each paragraph. In the conditional probability paragraph, one of two types of conditional probability information was presented: the probability of the evidence given the hypothesis, p(e/h), or the probability of the hypothesis given the evidence, p(h/e). The former is the expression of the reliability of the evidence that is pertinent input for Bayes' Theorem, which has been used in previous studies. The latter is equivalent to the requested answer. That is, the latter is p(h/e) ["the probability that hypothesis h is true, given that evidence e has been observed"], and the subject has been given evidence e, and is asked for p(h). The variation was accomplished by replacing several sentences; see the alternatives in Appendix II. The numbers were the same across these variations.

In the subject-selected order version, after responding to the introductory paragraph the subject was shown versions of the remaining paragraphs, from which the key information had been blanked out. The subject-selected order version of the Twins problem is presented in Appendix III. The subject selected the order in which he or she preferred to receive the (a) base rate, (b) evidence, (c) p(e/h), and (d) p(h/e) information. The presentation of the blank p(e/h) and p(h/e) paragraphs was counterbalanced between the 3rd and 4th positions, for each of the three problems. The information was subsequently presented (by filling in the blanks) in the requested order and the subject responded after receiving each piece of information. Identical values were used for the p(e/h) and p(h/e) information.

The values of the probabilities were as follows, for both the fixed order and the subject-selected order versions of the problems: Twins: p(h) = p(Stephen) = .20; p(e/h) = p(h/e) = .60; Doctor: p(h) = p(toxic uremia) = .15; p(e/h) = p(h/e) = .80; Insurance: p(h) = p(no accidents in next five years) = .35; p(e/h) = p(h/e) = .75.

2.3. Design.

Each subject was asked to do all three problems, with a fixed order problem in the first and third positions, and the subject-selected order problem in the second position. The Doctor, Insurance, and Twins problems were assigned to these three positions in all 6 possible orders [DIT, DTI, IDT, ITD, TDI, TID], between subjects.

Crossed with the identity of the problems, the conditional probability on the first problem was varied between p(e/h) and p(h/e). The conditional probability on the third problem was the opposite from that given on the first problem. Linked with this, the order in which the conditional probabilities were offered for the subject's consideration in the second problem was varied. To allow for the counterbalancing of problem and conditional probability, twelve different questionnaires were prepared.

2.4. Procedure.

The subjects were forewarned that they would be tape recorded. The researcher followed a detailed script (Appendix IV) which governed the explanation of the task, the explanation of and practice with thinking aloud, the turning on and off of the tape recorder, the delivery of reminders about the necessity of thinking aloud at prearranged points, and the asking of questions in the memory test. The order of problem presentation and the presentation of the conditional probabilities within the problems were determined by which questionnaire the researcher drew from the shuffled supply pile.

Subjects were run individually. A session started with explanation of the task and the thinking aloud procedure. For practice, the subject thought aloud while solving a mathematics problem. The first and third problems were fixed order problems. Subjects read each paragraph silently, then the tape recorder was turned on and they read the question aloud (see Appendix I) and answered it. Planned reminders to think aloud were given after the first and third paragraphs. The second problem was a subject-selected order problem. The first paragraph was done just as on the first problem. Then the researcher explained the procedure and the subject read all four blank paragraphs aloud, ordered them while thinking aloud, and answered the question after receiving each piece of information in turn in the order requested. One

point that the researcher made was that the subject should request information so as to get the right answer as soon as possible. The third problem was done as the first.

Following the completion of the third problem, the subject was unexpectedly asked four questions about the first problem. Because the study design varied the content of the first problem, across subjects, there are memory test data for all three problems (across subjects), and for subjects who had received both p(e/h) and p(h/e) information for each problem. The first question was open ended, "What was the first problem you did?" After the subject's answer, the researcher identified the problem and read the exact words of the question the subject had been asked in the problem. The other questions in the memory test were: "What information in the problem was pertinent for answering this question?" "What was your final answer?" and "How did you arrive at your final answer?" Subjects thought aloud and were tape recorded while responding to the memory questions.

2.5. Collection and analysis of verbal protocols.

The three goals for analyzing the verbal protocols were to determine (a) what information was attended to and used, (b) the strategies used for integrating the different kinds of information, and (c) whether the subject treated the conditional probabilities p(e/h) and p(h/e) differently. Protocols from thirty subjects (15 undergraduates, 12 math graduate students, and 3 insurance professionals) were analyzed.

2.5.1. The coding of the verbal protocols.

The procedure is described in detail in a "Coders' Materials and Reliabilities" document (Hamm, Lusk, Miller, Smith, and Young, 1988). Briefly, the tape recordings of subjects' thinking aloud during their sessions were typed as computer word processing files. Typists specified which location in the session all material came from. There were three problems plus the memory test. For the first and third problem, subjects thought aloud while answering after the first (introduction), second (base rate), third (evidence), and fourth (conditional probability) paragraphs. For the second problem, subjects thought aloud after the first (introduction) paragraph, while considering each of 4 blanked out paragraphs to decide in what order to have the blanks filled in [base rate, evidence, p(e/h), and p(h/e)], and while answering after receiving the information in each of those paragraphs. For the memory test, subjects thought aloud while answering the open ended question and probes about the important information, the final answer, and how they arrived at the final answer. The first author edited the transcripts to assure that the typist had labeled the material correctly and had not typed the subject's reading of the text. (That information could have given the coders information about what experimental condition the subject was in.) Subject identity codes disguised the subject type (undergraduate, graduate student, professional). Finally the transcripts were printed, with each page containing the subject's verbalizations at just one location (except for the memory test).

Four distinct coding schemes were developed and applied to the transcripts: one for the first and third problems, one for the second problem, one for the memory scheme, and one concerning the subjects' interpretations of conditional probabilities, which was applied at all locations. The first three coding schemes required the coder to make judgments about the subject's overall activity at a location. The fourth scheme required subjects to identify and categorize *each instance* of the use of a conditional probability concept.

2.5.2. Coding scheme for fixed-order problems.

The fixed-order problems are coded with respect to the information people attended to and the strategies they applied to that information in arriving at their answers. The codes are summarized in Table 1. After the second paragraph of each problem, which presented base rate information, coders looked for (1) the *mentioning of base rate*, (2) the *use of base rate*, and (3) strategies of *interpolation* or *anchoring and adjustment*. After the third paragraph of each problem, which presented evidence information, they looked for categories 1 and 2, plus (4) a revised set of *use of base rate* categories, (5) *mentioning of the evidence*, (6) *use of the evidence*, and (7) mention of the issue of the *reliability of evidence*. After the fourth paragraph of each problem, which presented conditional probability information, coders looked for categories 1, 3, 4, 5, and 6, plus (8) a revised set of *mention of reliability* categories, (9) *use of conditional*

probability, and (10) *use of Bayes' Theorem*. Coders' training materials for these coding schemes are presented on pages 4 to 9 of Coders' Materials and Reliability (Hamm, et al, 1988).

****
Insert Table 1 about here.
****

Reliability. Coder #1 coded all subjects and Coder #2 independently coded 4 randomly selected subjects. They used the same answer for 125 of the 136 coding groups (92%). More detailed reliability data are on page 20 of Coders' Materials and Reliability.

2.5.3. Coding scheme for subject-selected order problem.

The subject-selected order problem was coded for evidence concerning the subject's response to the two alternative descriptions of conditional probability information that were offered for their consideration (the third and fourth paragraphs in the list of possibilities). There were 4 possible categories:

- Level 0. The subject does not spontaneously mention any comparisons between the information in the two paragraphs.
- Level 1. The subject expresses some confusion or differentiation regarding the information in the two paragraphs, but does not resolve the differences or the confusion.
- Level 2. The subject expresses some confusion or differentiation regarding the information in the two paragraphs, and concludes that the information in the two paragraphs is the same.
- Level 3. The subject expresses some confusion or differentiation regarding the information in the two paragraphs, and concludes that the information in the two paragraphs is different.

Instructions for the coders are given on pages 8 and 9 of Coders' Materials and Reliability.

Reliability. Coder #1 analyzed all subjects. Coder #2 independently coded 4 randomly selected subjects, and agreed 100% with Coder #1.

2.5.4. Coding scheme for memory test.

The transcripts from the memory test were analyzed to determine whether and to what extent each of four concepts was used when the subject recalled the first problem, after each of four probe questions. The questions were: open ended recall, plus recall of important information, of final answer, and of how one arrived at one's final answer. The coder looked for four concepts after each question: base rate, evidence, reliability, and final answer. For each concept after each question, the coder made three responses:

1. If the subject mentioned the concept, mark 1; if no mention, mark 0.
2. If the subject stated the correct numerical value of a probability, or the correct evidence, mark 1; if incorrect, mark 0; if no mention, leave blank.
3. If the subject stated an incorrect value for a piece of information, write that value.

Coders' detailed instructions are on pages 9 to 11 of Coders' Materials and Reliability.

Reliability. Coder #1 analyzed all subjects, and Coder #2 independently coded four randomly selected subjects. The reliability of the coders' decision whether each of the 4 concepts was mentioned was measured. For the four subjects, there were 16 locations where each of the four judgments was made. The coders agreed on 15 of 16 locations for the base rate concept, on 13 for the evidence concept, and on all 16 locations for the conditional probability concept and the final answer. (More detailed analysis is presented on p 21 of Coders' Materials and Reliability.)

2.6. Verbal protocol coding scheme for conditional probabilities.

The final coding scheme focusses on the use of conditional probability concepts to express a relation between evidence and hypothesis. The purpose is to determine whether the specific expressions of these relations in the test of the word problems affected the expressions the subjects used. In the other three coding schemes, the coder made one judgment about everything the subject said at a particular location in the session. Here, the coder made a separate judgment about every instance of use of a conditional probability concept. For example, if the subject used conditional probability concepts 20 times after reading a particular paragraph, then the coder would make 20 separate category judgments. The number of times each category was used at a given location was recorded.

The conditional probability codes were applied at each of 9 locations in the transcript. Two locations were in the first (fixed-order) problem: the subjects' answers following paragraph 3 (evidence) and paragraph 4 (conditional probability). The same two sections were coded for the third problem. For the second (subject-selected order) problem, four sections were analyzed: while the subject decided in what order to receive paragraph 3 information (one of the conditional probabilities) and paragraph 4 information (the opposite conditional probability); and while the subject used the paragraph 3 information and the paragraph four information to produce an answer. The final location is the memory test, which was coded as a whole.

Coders did not know whether the conditional probability information given in paragraph four of the fixed-order problems (Problems 1 and 3) or recalled during the memory test was originally presented as p(e/h) or p(h/e). On the subject-selected order problem, they did not know which conditional probability was presented in paragraph 3 and which in paragraph 4.

Coders had to recognize that a conditional probability concept had been used, and then code it as one of seventeen categories (Table 2), which are defined for each of the three problems in Appendix V. There were four *specific inference* conditional probability categories: p(h/e), p(h/~e), p(~h/e), and p(~h/~e). The specific hypotheses (h and ~h) and piece of evidence (e and ~e) are defined for each problem in Table 3. Coders distinguished between expressions of p(h/e) that referred to the reliability information, and those that referred to the answer to the problem, and did not count the latter. There were four *specific evidence* conditional probabilities: p(e/h), p(e/~h), p(~e/h), and p(~e/~h).

There were two *general inference* conditional probability categories, p(correct inference) and p(wrong inference), and two *general evidence* categories, p(correct evidence) and p(wrong evidence). A statement was considered "general" if it expressed the probability of one class (evidence or hypothesis), conditional on the other class (hypothesis or evidence), without specifying the hypothesis or evidence. For example, "the probability the babysitter could misidentify one of the boys" is p(wrong evidence). Inference or evidence conditional probabilities were called "correct" if the evidence and hypothesis referred to the same event, and "wrong" if they referred to the complementary event. The transcripts were not coded for a general "p(correct)"; rather, such statements were called "ambiguous".

There were four *conjunction* categories: p(h and e), p(h and ~e), p(~h and e), and p(~h and ~e). [The transcripts were not coded for *general conjunctions*.] These were used when the subject mentioned a probability that involved both evidence and inference but did not specify a conditional relation between them. Use of this category addresses the possibility that subjects may interpret conditional probability statements as conjunctions (Pollatsek, Well, Konold, Hardiman, and Cobb, 1987). Finally, there was an *ambiguous* category for conditional probability statements where the subject could have intended two or more concepts of the other categories. If in the subject-selected order problem (Problem 2) it was not possible to determine whether the subject was referring to the information from paragraph 3 or paragraph 4, the coder excluded it from consideration. Coders' instructions for making these categorizations are given on pages 11 to 19 of Coders' Materials and Reliability.

Although the coders explicitly used 17 categories, there is an additional category to which most of the subjects' thoughts were assigned: those utterances that did not address the relation between evidence and hypothesis were deemed "irrelevant".

2.6.1. Combination of conditional probability categories into more general categorization schemes.

Coder 3 analyzed all subjects, and Coder 4 independently coded 10 subjects. The reliability of their coding will be analyzed in terms of their agreement (a) categorizing individual uses of the conditional probability concept, and (b) counting the uses of the concepts at given locations. The advantage of the first perspective on reliability is that its focus on individual responses clarifies the sources of disagreement between the coders. The experimental design, however, has manipulations that apply to locations (paragraphs in the problem), not to individual sentences in the transcript, and therefore the second perspective on reliability is needed.

The data are sparse. At each location, a coder could have identified a conditional probability category any number of times. However, most categories were not used in a given location for a given subject, and some of them never occurred. For example, neither coder used the category "p(wrong inference)" nor "p(~h and ~e)" on the transcripts of any of the 10 subjects in the reliability sample.

Most purposes of the study are still served if the categories are combined into more general categories that preserve the essential distinction between conditional probabilities that express the concepts of *evidence* [p(e/h)] and *inference* [p(h/e)]. Two levels of combination are used. These involve ignoring two of the following three distinctions that define our sixteen non-ambiguous categories: (*a*) whether the relation between the evidence and the hypothesis is (1) an *evidence* relation, e.g., p(e/h), (2) an *inference* relation, e.g., p(h/e), or (3) a *conjunction*, e.g., p(h and e); (*b*) whether the evidence and hypotheses are *consistent* (involving e and h, or ~e and ~h) or *inconsistent* (involving e and ~h, or ~e and h); and finally (*c*) the specific identity of the evidence and hypothesis, e versus ~e, and h versus ~h (e.g., in the Twins problem h is *Stephen* and ~h is *Paul*, as in Table 3). Ignoring distinction *c* provides the basis for our first combination of categories; ignoring both *b* and *c* is the basis for the second (see Table 4). The categories the coders assigned are in Column 1 of Table 4. When the specific identities of the hypotheses (*c* above) are ignored, the new categories still preserve information about the relation between evidence and hypothesis, and about whether these are consistent or inconsistent (Column 2). When the consistency information (*b* above) is also ignored, the categories reflect only the form of the conditional (or unconditional) relation between the evidence and the hypothesis (Column 3).

****

Insert Table 4 about here.

****

2.6.2. Reliability of categorization of individual uses of conditional probabilities.

Measures of the reliability of a coding scheme can be produced from a table comparing two coders' judgments, in the manner described by Rowe (1985) and Zwick (1988) (see Appendices II and III of Coders' Materials and Reliabilities). Table 5 shows both our coders' categorizations of the 71 uses of conditional probability concepts that were considered by both coders to be non-ambiguous out of all 9 transcript locations for all 10 subjects in the reliability sample. Over 95% of the phrases in these transcripts were not coded by either coder, and are not represented in the table. Also excluded are 88 instances that one coder coded as a non-ambiguous category but the other coder considered ambiguous or did not code. For the 71 concepts included in Table 5, coders assigned 47 (66%) to the same category and 24 to different categories.

****

Insert Table 5 about here.

****

A first indication of the amount of agreement between coders in Table 5 is provided by a test (X^2 = 360, df = 99, $p < .001$) showing significant dependency between their categorizations. However, this statistic is not an adequate measure of the agreement between two coders. Although the X^2 will be large when the coders assign instances to the same categories, it will also be large when there is some other form of dependence between them, such as reliably assigning the same type of instance to different categories. Zwick (1988) recommends using the π (Scott, 1955) or κ (Cohen, 1960) indices, but only if the two coders use the categories with similar relative frequencies. Zwick (1988) suggests that one is not justified in applying the π or κ indices unless a test comparing the coders' marginal distributions is not

statistically significant. Our data meet this criterion (X^2 = 10.5, df = 12, p < .50). The π is .591 and the κ is .593. These measures represent the proportion of agreement between coders, over and above what would have been expected by chance. They range from -1 to +1, where 0 means that there is only chance agreement between the coders.

Some of our analyses will use the combined categories defined in Table 4, so it is necessary to assess their reliabilities. If the categories in Table 5 are combined as indicated in the middle column of Table 4, a new table is produced (not shown; see pages 21-37 of the Coders' Materials and Reliabilities) whose measures are π = .573 and κ = .576. If the categories are further combined as in the rightmost row in Table 4, the new table has π = .578 and κ = .581. The reliability of the coding is very stable as the categories are combined.

These levels of agreement are lower than those found in a similar study by Rowe (1985, pp 166-180), where the κ's ranged from .812 to .962. However, Rowe coded all units (three-second segments of the tape) explicitly, while our coders excluded most units (about 95%) as irrelevant, before agreement statistics were calculated. Had the irrelevant units (phrases and sentences) been included in the analysis, they would have dramatically increased κ. By the same token, had we applied the statistic to all sentences that were coded as a specific category by *at least one* (instead of by *both*) of the coders, the κ would have fallen to .23.

After Coders 3 and 4 had coded their 10 subjects, differences were reconciled in discussions with the first author. Coder 3, who was more accurate, coded the remaining subjects.

2.6.3. Reliability of counts of conditional probabilities used at each location in the session.

For the purposes of the study, the subjects' use of the conditional probability concepts will be measured by the number of times each category was used at each location in the session. The reliability of this count can be measured by the correlation between the coders' counts of a category, over a number of locations (Table 6). For example, the correlation between the two coders' counts of instances of the category p(h/e), over the 9 locations in a subject's session, ranged from .15 to 1.0, with a mean (across subjects) of .70. The mean correlations (over the 10 subjects) for all the most specific categories (Column 1 of Table 4) are given in the top section of Table 6. The number of subjects for which both coders used the category in at least 1 of the 9 locations is given in the right-most column. Seven of the categories never met this criterion. For the most specific categorization scheme, the mean correlation for non-ambiguous categories ranged from .112 to 1.0. When the categories were combined (Column 2 of Table 4), the mean correlation (middle section of Table 6) ranged from .076 to .761. Finally, using the most general categorization scheme (Column 3 of Table 4; correlations presented in the bottom section of Table 6), the mean correlation between the coders' counts of the number of instances of *evidence* conditional probabilities [generically p(e/h)] was .42, and the mean correlation of their counts of *inference* conditional probabilities [generically p(h/e)] was .62.

****
Insert Table 6 about here.
****

3. Results.

The analysis of subjects' responses to the fixed-order problems will be presented first, followed by their preferences for key information in the subject-selected order problem, then their use of key information and of strategies, their memory for key information, and their interpretations of conditional probability information, all from the protocol analysis.

3.1. Responses in the fixed-order problems.

The subjects answered the question "What is p(h)?" after each of four paragraphs that (a) defined the problem, (b) gave base rate information, (c) gave evidence, and (d) gave a conditional probability expression of a relation involving evidence and hypothesis.

3.1.1. The use of available numbers.

It has often been observed that subjects respond to probabilistic inference word problems using numbers available in the text of the problem. The frequency with which available numbers were used in the present study is shown in Tables 7 through 9. Data from fixed-order problems presented in both the first and third positions are included in these tables. Hence each subject appeared in two columns, except for those who did only one or two problems. Many subjects used the base rate as the response after the paragraph in which it alone had been presented (Table 7), but it was seldom chosen after further information was received. Similar behavior was observed by Lyon and Slovic (1976), Hamm (1987a), and Ofir (1988).

****

Insert Table 7 about here.

****

The theory that people purposefully neglect statistical base rates because they do not pertain to unique cases would predict that the base rate would be attended more in the Twins problem, where it was drawn from experience with these twins, than in the other two problems, where it was drawn from experience with other people. However, a higher percentage of subjects used the base rate as the answer in the Doctor problem than the Twins problem.

The evidence information influenced subjects to have 100% belief in the hypothesis (hence completely ignoring the base rate information that they had previously been given) in 50% of the subjects in the Doctor problem, but in less than 20% for the other two problems. These differences reflect the context of the problems, before specific information about evidence reliability has been given: medical evidence is assumed to be reliable while identifications of twins and self-reports of accidents are assumed to be unreliable.

****

Insert Table 8 about here.

****

A high percent of the subjects use the conditional probabilities in the Doctor problem (Table 9), and the proportion was lower in the Twins problem (consistent with the results for these problems in Hamm, 1987a). No one used the conditional probabilities in the Insurance problem, which is new in this study. When they used the conditional probability, about equal proportions of subjects responded using the p(e/h) [*evidence*] and p(h/e) [*inference*] forms. Tests of whether different proportions of subjects use the different conditional probabilities are non-significant (Doctor problem: $X^2 = .07$, df = 1; Twins problem: $X^2 = .53$, df = 1). This supports the hypothesis that the two forms are confused.

****

Insert Table 9 about here.

****

3.1.2. Differences between first and third problems.

Subjects did two fixed-order problems. The experience doing the first and then doing the subject-selected order problem might have affected their strategies on the third problem. There were no differences in the mean answers for any problem when it was done in the first *versus* third position. The number of subjects who used the available base rate response after the second paragraph was not affected. There is a decrease in the number of subjects who give 100% as their answer after the third paragraph (when they have *base rate* and *evidence* information) in the Doctor problem, from 7 of 8 to 3 of 12. However, there were increases in the use of 100% on the Twins (from 0 of 15, to 3 of 9) and Insurance (from 0 of 10, to 4 of 12) problems. Overall, there was no change in the frequency of 100% responses. There may be a contrast effect here: experience with the Twins or Insurance evidence being unreliable may make people **not** expect that the evidence in the Doctor problem might be unreliable, while experience that the Doctor evidence is unreliable may make people expect that the evidence in the other problems will also be unreliable.

There was an increase in the number of subjects who used the conditional probability as their

answer after the 4th paragraph on the Doctor problem (from 4 of 8, to 11 of 12; $X^2 = 4.44$, df = 1, $p < .05$). Subjects continued to use p(e/h) as often as p(h/e).

3.1.3. Differences between subject types.

No differences were found between undergraduates and mathematics graduate students in the mean answers on the fixed-order problems, nor in their tendencies to answer with the base rate when only the base rate was available, with 100% when the base rate and evidence were available, or with the conditional probability when all three pieces of information were available. The insurance professionals' responses were similar to the other groups'.

3.2. Order in which information was requested in subject-selected order problems.

Because subjects were instructed to try to get accurate answers as soon as they could, the order in which they requested the information should reflect their judgment of the information's usefulness. The number of subjects who requested that the *base rate*, *evidence*, *p(e/h)*, and *p(h/e)* information be revealed to them in each possible order is presented in Table 10. Fifteen subjects (47%) requested the base rate information first, which is inconsistent with the Purposeful Neglect of Base Rate Hypothesis. Eighteen subjects (56%) wanted both base rate and evidence information before either of the conditional probabilities. This is consistent with the Interpolation Hypothesis, where it would be sensible to learn the numbers one must interpolate between, before one learns the number governing the interpolation (the conditional probability). (Note that the base rate and evidence were always listed in the first and second positions on the questionnaire, respectively, so we can not test for what effect their position early in the list may have had on their being requested by people who had no strong preferences.) Eight subjects (25%) asked for both evidence and a conditional probability before they asked for base rate [6 of them asked for **both** conditional probabilities, one just for p(h/e), and one just for p(e/h), before the base rate]. This is consistent with the Confusion Hypothesis, that is, with the belief that the conditional probability is an appropriate answer when the evidence is known. Seven subjects chose the evidence **after** the conditional probabilities (expressing opinions such as that the evidence is useless unless one knows its reliability). Apparently these subjects felt they needed all information before they could answer. This is not consistent with any of the three theories. It would be appropriate only if one thought one needed all the other information before one could interpret any of the information.

****
Insert Table 10 about here.
****

The preference for base rate *versus* evidence depends on both the problem and the type of subject. Table 11 shows the ordinal position in which base rate and evidence were selected, separately for each problem. The number of times base rate was selected at each position differed between problems ($X^2 = 10.53$, df = 6, $p < .15$), it being selected in the first position more often in the Insurance and Twins problems, than in the Doctor problem. The evidence information, on the other hand, was selected first more often in the Doctor problem ($X^2 = 14.35$, df = 6, $p < .03$). This is related to Cohen's (1981) claim that people find statistical base rates more relevant when they pertain to the particular case at issue. In the Twins problem, the base rates were pertinent for these particular twins, while in the Doctor problem and the Insurance problem, the rates pertained to people in general. Subjects requested base rate information earlier on the Twins problem than on the Doctor problem, as Cohen would predict, but the Insurance problem is anomalous.

****
Insert Table 11 about here.
****

Table 12 shows the subject types' preferences for base rate and evidence. The overall pattern is not significant, but there is a suggestion that the graduate students prefer the base rate while the undergraduates prefer the evidence. This is consistent with Cohen's (1981) argument that correct answers (using base rate) on the probabilistic inference word problems are signs of people's education rather than of their common sense. The pertinent analysis compares the number of cases in which base

rate is selected before or after evidence for the two groups. Undergraduates preferred evidence (9 wanted it before base rate, and 6 wanted it after) while mathematics graduate students preferred base rate (9 wanted it before evidence, and 5 after), but the pattern is not significant (X^2 1.71, df = 1, $p < .20$).

****
Insert Table 12 about here.
****

3.2.1. Preference for conditional probability information.

Inspection of Table 10 shows that 30 of the 32 subjects requested the two conditional probabilities in adjacent positions and 18 wanted them last (in positions 3 and 4). This suggests that the subjects did not value the conditional probability information, or did not know what to do with it. However, more specific conclusions can be derived from a detailed analysis of the possible distinctions subjects may make between p(e/h) and p(h/e), and the possible orders in which it would be sensible to request information if one made these distinctions. Here are four possible types of distinction:

1. The subjects may know exactly what they will do with each piece of conditional probability information, and request them in a *purposeful* order.
2. The subjects may want *both* pieces of conditional probability information, and not be able to make any answers until they have them both (and possibly the other information too).
3. The subjects may consider the two conditional probabilities to have *identical* meaning.
4. The subjects may be aware that they are *confused* about the meaning of the two conditional probabilities.

It would be rational for subjects who make each of these types of distinction between the conditional probabilities to adopt different strategies for acquiring the p(e/h) and p(h/e) information in the context of the base rate and evidence. Assuming the subject applies these strategies without error, we can work backwards from the observed information request orders to infer the types of distinction the subjects must have made. Our analysis of these distinctions is summarized in Table 13.

****
Insert Table 13 about here.
****

If the subjects know exactly what they want to do with the information (*purposeful* type), then any order is possible (Column 2 of Table 13), depending on the subjects' plan. Knowing a subject makes purposeful distinctions between the conditional probabilities allows us to make no predictions about the order in which subjects will request information; but if they make such distinctions, the order in which they request the conditional probabilities informs us which one they think is more important (Column 3). If the subjects believe they can do nothing with conditional probability information until they have both conditional probabilities (*both needed* type), then the conditional probabilities will be requested in adjacent positions (Column 2), although the order in which they are requested will be arbitrary. Therefore, the orders will tell us nothing about the relative importance of the two conditional probabilities (Column 3). If the subject thinks the two conditional probabilities mean the same thing and hence are mutually substitutable (*identical* type), then any order is possible, but the second conditional probability is redundant and can be requested last (Column 2). Finally, if the subjects are confused about the meaning of the conditional probabilities and know it (*confused* type), then when they want one, they will have to ask for both to assure that they get the needed information. The two conditional probabilities will be asked for in adjacent positions (Column 2), in arbitrary order (Column 4).

Table 14 displays which of these types of distinction between conditional probabilities each of the subjects' information preference orders are consistent with. The first column shows the number of subjects who requested the information in each order. The other columns each represent a type of ordinal relation between the conditional probabilities p(e/h) and p(h/e). The possible types of distinction between conditional probabilities that each of these types of ordinal relation is compatible with are also listed as part of the column heading. If a particular order (row of Table 14) has a given relation between conditional probabilities, then the subjects who requested that order are counted in the relation's column. The column

sums are the number of subjects who had each conditional probability relation, and they can be compared with the number of subjects who would have the relation if all requests were random (bottom row). The sums of all columns that are consistent with a given type of distinction between conditional probabilities provide a measure of how likely it is that subjects make that type of distinction.

****
Insert Table 14 about here.
****

Consider first the possibility that subjects might know in exactly what order they need the conditional probability information (*purposeful* distinction type). If so, the relative positions of the conditional probabilities would reveal their relative importance. The last two columns of Table 14 show the number of subjects who picked p(e/h) before, or after, p(h/e). Twenty of the subjects requested p(h/e) before p(e/h) and the other 12 requested p(e/h) first, a difference which is not statistically significant ($X^2 = 2.0$, df = 1, $p < .25$). If subjects know which conditional probability they want, then they do not all want the same one.

If subjects make a distinction between the conditional probabilities, prefer one, and request it before the other, then this preference should be stable over minor variations in the problem presentation (see Column 4 of Table 14). If, however, they consider the conditional probabilities to be *identical*, *confusing*, or *both needed*, then they should have no clear preference, and the order in which they request the conditional probabilities may be influenced by other factors, such as the order in which the conditional probability information was listed when offered to the subjects for ordering. This was an independent variable in the study, counterbalanced within problem (Doctor, Insurance, Twins). Table 15 shows the number of people who requested the conditional probabilities in each possible order, as a function of the presentation order. Subjects selected the first-listed conditional probability 22 times, and the second one 10 times ($X^2 = 4.10$, df = 1, $p < .05$). This implies that they do not prefer one conditional probability over the other, and therefore that they do not make a *purposeful* type of distinction between the conditional probabilities.

****
Insert Table 15 about here.
****

For the remaining types of distinction in Table 13, there is no reason for the subject to request one of the conditional probabilities before the other. However, the remaining types of distinction would involve different information seeking strategies: if subjects think the two conditional probabilities are identical, they will ask for one of them last; if subjects *need both* conditional probabilities before they can use either, or know that they are *confused* between them, then they will request them in adjacent positions. Comparisons among the distinctions are represented in Columns 2, 3, and 4 of Table 14. Column 2 represents those preference orders in which the conditional probabilities were requested in non-adjacent positions, with one in last position. This is consistent only with the subject considering the two conditional probabilities to be *identical*. (No one requested the conditional probabilities in non-adjacent positions with neither in the final position. Such an order would be consistent only with a *purposeful* distinction type.) Column 4 represents those preference orders in which the conditional probabilities are requested in adjacent positions, but not at the end. It is consistent with subjects *needing both* conditional probabilities, or being *confused* about them, but not with thinking they are *identical*. When the conditional probabilities are requested in the last two positions, it is compatible with all three of these types of distinction (Column 3). Each subject is placed in one of Columns 2, 3, or 4.

Twenty of the 32 subjects chose the information in an order that was compatible with believing that the conditional probabilities are *identical*, in comparison with a random expectation of 16 ($X^2 = 2.0$, NS). Thirty of the 32 had an information preference order consistent with either *needing both* conditional probabilities, or being *confused* about them and asking for both to assure they get the information they need ($X^2 = 22.78$, df = 1, $p < .001$). (Note that any of these orders is also compatible with knowing exactly what one wants to do with the conditional probabilities. This possibility has been excluded due to the effect of the order in which the conditional probabilities are presented, discussed above.) Thus there is more support for the *confused* and *both needed* distinctions than the *identical* one. Note also that very few subjects chose the preference orders that were consistent with only the *identical* distinction.

In summary, analysis of the order in which subjects requested information has shown: (a) subjects tended to prefer the base rate and evidence information over the conditional probability information; (b) their preference for p(e/h) and p(h/e) was influenced by the order in which the conditional probabilities were offered; and (c) they tended to request the two conditional probabilities in adjacent positions. The first finding is inconsistent with the Purposeful Neglect of Base Rate Hypothesis, and the second and third finding are consistent with subjects being unable to distinguish the two conditional probabilities (the Confusion Hypothesis) but also with their needing both conditional probabilities (reasonable under the Interpolation Hypothesis).

3.3. Responses in subject-selected order problems.

The distributions of final answers to the second problem are similar to those of the fixed order problems, which suggests that selecting the order to receive the information does not change people's answers.

Subjects' answers following the receipt of conditional probability information should be related to their interpretations of the conditional probabilities. Even though identical numerical values were given for the two conditional probabilities, the answers should be different, though the exact nature of the difference depends on the other information available. However, twenty six subjects gave identical answers following receipt of the p(e/h) and p(h/e) information, only 4 gave larger answers after p(e/h), and 3 gave larger answers after p(h/e).

3.4. Protocol analysis: Attention to and use of different types of information.

The protocol analysis data allow us to study what information the subjects pay attention to and how they use it in producing their answers. This analysis is done for the fixed-order problems, where the *base rate* information was presented in the second paragraph, the *evidence* in the third paragraph, and the conditional probability information [either *p(e/h)* or *p(h/e)*] in the fourth paragraph.

3.4.1. Subjects' attention to and use of base rate information.

To what extent do people neglect base rate information? How much attention do they pay to the base rate information when it alone is given (paragraph 2) and when other information has been given subsequent to it (paragraphs 3 and 4)? Is it forgotten as new information comes in? Table 16 shows the number of subjects who were coded in each category of the coding schemes concerning (a) attention to base rate and (b) use of base rate. After receiving the base rate information, subjects mentioned the base rate on 55 of the 59 problems (each subject is represented twice in this table, for the first and the third problem, except for one who did not do the third problem), and used it as the answer on 40 of them. On 9 problems they used it in an adjustment process, either starting with it and adjusting, or starting with another number and using the base rate as a justification for the direction or the extent of adjustment. Seven of those who mentioned the base rate information did not use it, which shows that mentioning the base rate does not guarantee that it will have a discernable role in the production of the answer, even when it is the only specific information given.

****

Insert Table 16 about here.

****

When the evidence information was given in the third paragraph, the subject was confronted with competing case and base rate information. The correct answer could be anywhere between the base rate and 100%, depending on the subject's assessment of the reliability of the evidence (Hamm, 1987a). Subjects mentioned base rate on 21 of the 59 problems, used it as their answer on 9, and used it in an adjustment process on 9. More people attended and used the base rate in the Insurance and Twins problems than in the Doctor problem (mentions: $X^2 = 8.64$, df = 1, $p < .01$; uses: $X^2 = 8.97$, df = 1, $p < .01$; with the reservation that each subject appeared twice in the table). This is probably due to their prior expectations concerning the reliability of evidence in these problems.

There were no subject type differences in the distribution of problems over categories after the second or the third paragraph (not shown). Subjects payed more attention to the base rate after receiving the evidence (paragraph 3) on the first problem than they did on the third one.

The fourth paragraph presents conditional probability information that expresses the reliability of the evidence. On 31 of 58 problems, subjects mentioned the base rate. They used it as their answer on 6 of them (justifying this with reference to the unreliability of the evidence on 4), used it in conjunction with other information on 17, and did not use it at all on 33 problems. As with paragraph 3, after paragraph 4 subjects used the base rate information less on the Doctor problem than on the other two ($X^2 = 7.59$, $p < .01$; with the above reservation).

Undergraduates had a different approach from graduate students. Eighteen of 24 graduate students mentioned base rate, but only 11 of 28 undergraduates did ($X^2 = 6.68$, $df = 1$, $p < .01$; with the above reservation). One mathematics graduate student used the base rate as the response, 12 used it in conjunction with other information, and 8 did not use it. In comparison, 4 undergraduates used the base rate as the response, 4 used it in conjunction with other information, and 21 did not use it. The difference in these distributions is significant ($X^2 = 10.62$, $df = 2$, $p < .01$). This suggests the mathematics graduate students are more cognitively complex, using base rate in conjunction with other information, while the undergraduates are more likely to either ignore it or use it alone as their response. Thus the graduate students are more consistent with the 5th qualitative Bayesian principle.

Significantly fewer subjects used the base rate in any way after paragraph 4 on the third problem (4 of 24) than the first problem (14 of 27; $X^2 = 6.89$, $df = 1$, $p < .01$).

Does mentioning the base rate mean that it influences the subject's answer? And did anyone who was not observed to *mention* the concept, use it nonetheless? The relation between subjects' verbalizations and their answers is addressed by the data in Table 17. The base rate is always a low probability, and the mean answer for those who mention it was lower than for those who did not in all 8 conditions where the comparison could be made. Only four subjects (of 66) used the base rate as a response when they had not mentioned it, while 45 of 110 used it when they had mentioned it. This establishes a connection between the content of the subjects' thinking and their responses to the questions.

****
Insert Table 17 about here.
****

3.4.2. Subjects' attention to and use of evidence information.

Subjects were given evidence in paragraph 3 of the fixed-order problems, and of course it was still available for their use after they received the fourth paragraph. We have discussed above (Section 3.3.1) how frequently the subjects respond using 100%. The present analysis adds information concerning their attention to the evidence and their use of it in combination with other information. After receiving the evidence information, nearly everyone mentions the correct evidence (upper section of Table 18). Most use it on the Doctor and Twins problems, though only half do on the Insurance problem. Among those who use it, most "buy it 100%" on the Doctor problem, but most use it in an adjustment process on the Twins and Insurance problems.

****
Insert Table 18 about here.
****

When they got the reliability information, only about half of the Doctor and Insurance subjects explicitly mentioned the evidence while producing their answer, although 80% mentioned it on the Twins problem. This can not mean that they were not affected by the evidence. They knew, of course, which evidence had been delivered. But there was little trace in the transcripts of an explicit role for this knowledge. That is, most subjects did not say they had 100% belief in the hypothesis because of the evidence, nor did they talk about the evidence in the process of arriving at their answer. Even so, the

evidence is crucial in the production of the answer. For example, those who use the conditional probability p(h/e) or p(e/h) as their answer, e.g., .80, would have used its complement, .20, had the opposite evidence been observed.

As with the base rate results, there was a difference between undergraduates and mathematics graduate students (data not shown): the graduate students were more likely to use the evidence in conjunction with other information, while the undergraduates did not refer to it while producing their answer.

The relation between subjects' mentioning of evidence and their answers (data not shown) is weaker than the relation between mentioning of base rate and answers (in Table 17).

3.4.3. Subjects' attention to and use of conditional probability information.

Do subjects spontaneously consider whether the evidence is reliable? The upper part of Table 19 presents the results of the conditional probability coding after paragraph 3, where evidence, without explicit reliability information, was presented. Surprisingly, more subjects considered the reliability of the evidence on the Doctor problem than the other two (comparing the first 2 against the last 4 columns, dropping the middle row, $X^2 = 7.08$, df = 1, $p < .01$). It was expected that more people would spontaneously consider the reliability of the evidence after the third problem, because they had been forced to consider it on the first two problems. Seven of 28 subjects talked about it after the first problem (data not shown), and 11 of 27 after the third, but this increase is not significant ($X^2 = 1.54$, df = 1). The proportion of mathematics graduate students who considered the reliability of the evidence was not different from the proportion of undergraduates.

****

Insert Table 19 about here.

****

Subjects in earlier studies used the conditional probability p(e/h) information as their response on probabilistic inference word problems, although in this study that happened frequently only with the doctor problem. In paragraph 4 we presented either a p(e/h) or a p(h/e) expression of reliability, and we test whether subjects thought differently about the alternative conditional probabilities. Comparing the p(e/h) and p(h/e) columns in the second section of Table 19 for each problem shows that the type of conditional probability had little effect on the subjects' tendency to use the conditional probability as the answer, to use it in conjunction with other information, or not to use it. We saw above (Section 3.1.1) that more subjects used the conditional probability as their response on the doctor problem than the others. Table 19 reveals more about this: half the subjects on the other two problems *did* use the conditional probability, in conjunction with other information, while the other half did not reveal any use of it. Those who used it are behaving in accord with the 5th qualitative Bayesian principle.

Undergraduates are likely to use the reliability information as the answer, mathematics graduate students to use it to adjust their answer, and professionals to not use it (Table 20). This pattern is statistically significant ($X^2 = 15.51$, df = 4, $p < .01$), with the reservation that each subject is counted twice in this table and so the observations are not all independent.

****

Insert Table 20 about here.

****

Are the subjects' responses related to the extent they talked about the reliability of the evidence? Most subjects mentioned the conditional probability information after the 4th paragraph, and their answers were not systematically different from the answers of those who did not.

3.4.4. Subjects' response to having two conditional probability options on subject-selected order problem.

In the second problem, subjects were offered both p(e/h) and p(h/e) information. Their verbalizations while deciding the order in which to request the information, and while answering after receiving the information, were coded into four categories. Thirty subjects' ordering verbalizations were coded. Twenty-five commented on the two conditional probabilities. Twelve of these decided that the two had the same meaning, 9 decided they were different, and 4 expressed confusion which they did not resolve. Thus only 36% of those who talked about it concluded there was a difference between p(e/h) and p(h/e).

3.5. Memory for information from the first problem.

The analysis of the subjects' memory for the first problem is similar to the previous analysis. Table 21 shows the number of subjects mentioning each concept after each question. Separate analyses (data not shown) revealed no differences between problems nor between subject types, so this table collapses across those variables. Subjects either (a) did not mention the concept, (b) mentioned the concept without a specific number, (c) mentioned it and gave a wrong number, or (d) mentioned it with the right number. The evidence had no number, so only the first two categories could be applied. There are fewer subjects for the first question because 5 of them generated no verbalizations until the researcher had given information about the question.

****
Insert Table 21 about here.
****

At each question, the number of subjects recalling the base rate is comparable to the number of subjects recalling the evidence. The biggest difference is after the "key information" question, where 20 of 28 subjects recalled the base rate and only 15 of them recalled the evidence ($X^2 = 1.90$, df = 1, $p < .25$). The Purposeful Neglect of Base Rate Hypothesis would have predicted better recall for the evidence since it is pertinent while base rate is not. The similar recall patterns in the Twins problem (where the base rate is particular to *these* twins, and so the theory would say people should recall it) and the other two problems is also incompatible with this hypothesis (in parallel with the similar use of base rate on the fixed order problems; see Section *** above).

There is a (non-significant) hint of more frequent recall of conditional probability information when p(h/e) was originally presented. Six subjects spontaneously recalled p(h/e) on the open ended question, while only 2 subjects spontaneously recalled p(e/h). Increases in the number of subjects recalling conditional probabilities on the later questions also favored p(h/e). After the final question, 10 of the 15 subjects who had received p(h/e) mentioned the concept, while only 5 of the 13 who had received p(e/h) did ($X^2 = 2.23$, df = 1, $p < .25$). This might reflect a judgment that p(h/e) is more pertinent to the problem, and may be related to subjects' requesting p(h/e) before p(e/h) more frequently on the second problem.

3.6. Protocol analysis: Subjects' strategies for integrating information on fixed-order problems.

Coders sought evidence for three specific strategies for combining information: *anchoring and adjusting*, for example, starting at the base rate and adjusting upwards when given evidence; *interpolation*, as in starting with the base rate (at the low end), and the 100% expressing complete faith in the evidence (at the high end), and interpolating between them; and *Bayes' Theorem*, which is in effect a particular method for using the reliability information to guide an interpolation between the base rate and 100%.

If on paragraphs 2, 3, or 4 the given information was used "in conjunction with some other information", then the coder attempted to discover what kind of strategy was used (see Coding sets 3 and 10 in Table 1, and Coders' Materials and Reliabilities). For most subjects who did not reply using one of the pieces of information directly, it was not possible to categorize their process as using one of these three strategies. On paragraph 2, where only base rate was available, on only 9 of 59 problems (each subject did 2 problems) did subjects anchor and adjust, and none interpolated. There was more anchoring

and adjustment on the insurance problem. On paragraph 3, where base rate and evidence were available, subjects anchored and adjusted on 11 of 59 problems, and interpolated on 2. On paragraph 4, they anchored and adjusted on 7 of 59 and interpolated on 5. These data are too sparse for comparisons between problems and subject types. Subjects used **one** anchor more frequently than they used **two** anchors.

An analysis was done to determine whether subjects mentioned the same information or used the same strategy on the first and third problems. There was no individual stability in the mention or use of base rate. The data are too sparse for statistical tests (data are available from author). Of the 9 people who used either an interpolation or an anchoring and adjustment strategy on the 3rd paragraph on at least one of the two problems, 4 used the same strategy on both problems. There was, however, no such stability on the 4th paragraph.

3.6.1. Use of Bayes' Theorem.

Only four of the 29 subjects, all mathematics graduate students, attempted to use Bayes' Theorem on the fixed-order problems. Three of them used it on both fixed-order problems, and the other used it only on the last problem. Their use of Bayes' Theorem was not correct. They were coded as using it if they multiplied the probability of a hypothesis times a conditional probability expression of reliability, or if they said they were using it (see Coders' Materials and Reliabilities). For example,

> "I got the answer by assuming she was right 60% of the time in identifying them, but given the twins' track records of Paul usually the trouble..., I multiplied those two together, and got..., I multiplied the fact..., the chance that Stephen actually caused the trouble was 20%, the chance that she saw him doing it was 60%, so I multiplied them together and got 12."

In symbols, p(h) = .20, and p(e/h) = .60, so p(h)*p(e/h) = .12. The subject is neglecting the possibility of misidentifying Paul, p(~h) = .80 and p(e/~h) = .40. This is typical of most of the subjects' attempts to apply Bayes' Theorem. This neglect of the possibility of a false alarm confirms findings of Doherty et al (1979) and Beyth-Marom and Fischhoff (1983).

One subject applied Bayes' Theorem to the p(h/e), p(h), and evidence information, and got the answer that would have been correct if the p(e/h) information had been given. That is, the only subject of 32 in this study who solved one of the fixed-order problems by applying the mechanics of Bayes' Theorem correctly, did so after having misconceived the given p(h/e) as p(e/h). This is an exquisite example of the confusion that is at the heart of people's performance on these problems.

It would be interesting to ask whether the answers produced using the anchoring and adjustment or interpolation strategies are any different, and any more accurate, than those produced using other methods. However, these strategies were used so rarely that the comparisons have little meaning; none were statistically significant.

3.7. Protocol analysis: Subjects' interpretations of conditional probability information.

The following excerpt from a subject's transcript shows why it is difficult to figure out the subject's interpretation of the conditional probability, and the role it plays in the subject's decision whether and how to use the base rate information. This subject is working on the Doctor problem, and has previously received the base rate (15% of people with this symptom have toxic uremia) and the evidence (the Spock test says this patient has toxic uremia). The most recent information is the conditional probability, which in this case is p(e/h): the probability that the Spock test will say toxic uremia, given that the patient has toxic uremia, is .80.

> S: Ok. With what you now know - with what you know now, what is the probability that the patient has toxic uremia? So the problem is what disease does the guy have. And the information that I have is that he's given the Spock test, that is not perfect, and that it's right

> 80% of the time **[p(correct) = .80]**, when the patient has toxic uremia, the Spock test says so 80% of the time **[p(e/h) = .80]**. When he has hepatitis, it says that he has toxic uremia 20% of the time **[p(e/~h) = .20]**. Now, given that there are - he is ill, so he has one disease or the other, given that, you can't have both diseases, and that there are only those two, then, whether he has hepatitis or toxic uremia, this test is going to say that he has... The test - there's an 80% chance that the test was correct **[p(correct)]**, and that he does have toxic uremia **[p(h/e)]**. It seems to me, well... that the earlier information about... the fact that he had... 85% of people with symptoms of yellow eyes have hepatitis, and only 15% have toxic uremia **[base rate]**, doesn't matter any more, because, the fact is that, whether he has hepatitis or toxic uremia, the test is going to say so 80% of the time, correctly, which one it is **[p(correct)]**, so, um, I'd have to say the probability of his having toxic uremia is 80%.

Tracing the subject's interpretation of the conditional probabilities (**bold** text in the quote), it can be seen that the subject does not go directly from p(e/h) to p(h/e), but rather, understands p(e/h), generalizes to p(correct), and then applies p(correct) to produce p(h/e). This is also an example of a decision to neglect base rate following the conditional probability.

In the next example, a subject acts as if the two conditional probabilities are either the same thing, or have the same value. The subject has just received a conditional probability (the fourth paragraph, in the subject-selected order problem):

> So, now we say, if it is Stephen, you think it's Paul two times in five **[p(~e/h) = .40]**. But she thought it was Stephen, so if it is Paul..., so two times out of five, if it is Paul, you think it's Stephen **[p(e/~h) = .40]**. She thought it was Stephen **[evidence]**, <inaudible> two times in five it's going to be Paul **[p(~h/e) = .40]**, so the chances are <inaudible>. I guess, I'm going to say... chances are lower though, for Stephen, I would give him, 60% chance of breaking it **[p(h/e), the required answer]**.

This subject states "evidence" conditionals, and then, in using them, switches to an "inference" conditional, without recognizing that anything unusual has been done.

3.8. Frequency of various interpretations of the conditional probability information.

The particular concepts the subjects used when discussing the reliability of the evidence were coded (see Section 2.6). The concepts were not used equally frequently (Table 22). On the average, during a session subjects made 6.48 statements about *evidence* conditional probabilities [p(e/h), p(~e/h), p(e/~h), and p(~e/~h)], 4.82 about *inference* conditional probabilities [p(h/e), p(~h/e), p(h/~e), and p(~h/~e)], 0.37 about conjunctions, and 10.22 statements that were ambiguous. This means that there were few of each type of conditional probability statement at each of the 9 locations in a session. To compensate for the relatively low counts for the specific categories, we shall report analyses using the most general categorization scheme, described in Section 2.6.1. Although the frequent use of ambiguous statements is a sign of imprecise interpretation of the conditional probabilities, it is not direct evidence for misinterpretation and so is excluded from the following analysis.

****

Insert Table 22 about here.

****

3.8.1. Subjects' responsiveness to the conditional probability information on the fixed-order problems.

Do subjects use the particular conditional probabilities that they are given in the text of the word problem, when deliberating about the reliability of the evidence? Table 23 shows the pertinent results for the fixed-order problems, using collapsed categories of *evidence* and *inference* conditional probabilities. Table 24 shows the analogous results for the subject-selected order problem. Table 25 summarizes these and also shows the results for the memory test. The upper half of Table 23 represents those subjects whose reliability information (in paragraph 4) was presented as p(e/h) on the first problem, and as p(h/e)

on the third problem. The lower half shows subjects who received the conditional probabilities in the reverse order. If the given conditional probability expressed the *evidence* concept, p(e/h), as in the upper left and lower right cells of Table 23, then the conditional probability concepts that the coder thought the subject was using should have been *evidence* concepts. Conversely, if the given conditional probability expressed the *inference* concept, p(h/e), then the conditional probability concepts the subjects used when working on answering the question should have been *inference* conditional probabilities. (Some possible uses of the non-presented conditional probability are appropriate, of course, but usually they were due to a misinterpretation, as in the first transcript in Section 3.7.)

****
Insert Tables 23 and 24 about here.
****

The counts in Table 23 were analyzed with a repeated measures analysis of variance. (The responses of the 3 professional subjects were excluded.) The key hypothesis is embodied in the 3-way interaction of (a) the order in which the conditional probabilities were given [p(e/h) on the first problem and p(h/e) on the third, or *vice versa*] by (b) first or third problem by (c) the type of conditional probability the subject talked about [*evidence* or *inference*]. Overall, the cells which represent subjects' use of the conditional probability concept that was presented have a larger count than the cells representing the opposite conditional probability. This was marginally significant ($F(1,22) = 3.45$, $p = .077$). Main effects and interactions involving the type of subject were not significant, although undergraduates seemed to use these concepts less than mathematics graduate students ($F(1,22) = 2.02$, $p = .17$).

Results ignoring subject type and the distinction between first and third problem are displayed in the top section of Table 25. When *evidence* conditional probabilities were presented, 59% of the specifically identifiable conditional probability concepts the subjects used were of the p(e/h) form, and 41% were of the p(h/e) form. When *inference* conditional probabilities were presented, 62% of the identifiable concepts were appropriate. Overall, 60.3% of the concepts the subjects used were the ones they had been presented with, and 39.7% were the opposite. This pattern occurs equally for the doctor, twins, and insurance problems.

****
Insert Table 25 about here.
****

The analysis of Table 23 contrasted the *evidence* and *inference* types of conditional probabilities at their most general. When the orthogonal dimensions of (a) whether these relations were between hypothesis and evidence concerning the same hypothesis or the complementary one and (b) the identity of the particular hypotheses (e.g., "Paul did it" or "Steve did it", in the Twins problem; see Table 14) were added to the analysis, they did not themselves have significant effects, and the key pattern was unaffected.

3.8.2. Subjects' responsiveness to the conditional probabilities on the subject-selected order problems.

The second problem offers a separate opportunity to measure the extent to which the conditional probability expressions the subjects used when talking about the reliability of the evidence were the same form as what was given to them. Here, subjects were shown both of the conditional probabilities, as the third and fourth options in the list of information that they could request. Data from their discussion of these paragraphs while deciding in what order to receive the information, and while answering the problem after having received the information, are combined for this analysis. The mean number of times each subject used *evidence* and *inference* type conditional probabilities, when the reference of the concept could be traced to a specific paragraph, are shown in Table 24.

Subjects mentioned specific conditional probability concepts (both *evidence* and *inference*) less often when they were deciding in which order to receive the information (mean = 1.81 mentions, for 26 subjects) than when they were answering the question following receipt of the information (mean = 4.35 mentions; data not shown). Neither the paragraph (third or fourth) nor the order in which those

paragraphs contained the p(e/h) and p(h/e) conditional probability concepts [p(e/h) in the third paragraph and p(h/e) in the fourth, or vice versa], nor their interaction, had an effect on the number of conditional probabilities mentioned. Subjects tended to mention *evidence* type conditional probabilities more often (mean = 3.73, summing over their paragraph 3 and paragraph 4 mentions for both ordering and answering) than they mentioned *inference* type conditional probabilities (mean = 2.42; $F(1,24) = 3.87$, $p = .061$). There was no difference in the numbers of conditional probability concepts that were elicited by *evidence* versus *inference* conditional probabilities (mean = 3.38 concepts when talking about the *evidence* conditional probabilities versus 2.78 when talking about *inference* conditional probabilities, $F(1,24) = 0.62$, $p = .442$).

The key hypothesis, that subjects use the presented conditional probability concept more often than its opposite, is tested by the 3-way interaction of (a) the stimulus paragraph the subject was talking about (third or fourth) by (b) which of these paragraphs had the *evidence* and *inference* conditional probabilities (3rd = p(e/h) and 4th = p(h/e), or vice versa) by (c) the type of conditional probability the subject talked about (*inference* or *evidence*). Generally, subjects used the conditional probability that had been presented more often (4.44 times per subject) than its opposite (1.96 times; $F(1,24) = 11.67$, $p = .002$). This pattern was not affected by subject type, problem content, or subject's choice of whether to get the p(e/h) or p(h/e) information first. To put it in perspective, 69.4% of the unambiguous conditional probability concepts were appropriate, and 30.6% were inappropriate (see middle section of Table 25).

The use of the presented conditional probabilities, embodied in the three-way interaction pattern starred in Table 24, was different when the subjects were reading the paragraphs to decide in what order to get their blanks filled in, than when they were reading the paragraphs and trying to answer the original question using the information that had now been given (data not shown). Subjects tended to use the presented conditional probabilities more when answering the problem, than when deciding in what order to request the information.

Subjects proved able to discriminate between the conditional probability expressions of reliability on this problem, in that the concepts they used were significantly related to the concepts on the page before them. They did so more on these subject-selected order problems than on the fixed-order problems, which may reflect the fact that the presentation of both conditional probabilities called attention to the interpretation of the conditional probabilities. The finding that there was a higher proportion of accurate use of the presented conditional probability when subjects were answering the problems, which presumably requires careful thought, than when selecting the order in which to receive the information, offers additional support for this "depth of processing" explanation.

Although subjects discriminate between the two types of conditional probability when thinking aloud, they use the same answers following the two conditional probabilities. This suggests that the distinctions they make here may not be a stable part of their understanding. Their verbalizations may be controlled by the stimulus more than are the concepts they use. This is tested by their recall -- do they distinguish as accurately on the memory test as they did on the original problem?

3.8.3. Subjects' responsiveness to the conditional probabilities when recalling the first problem.

The third test of whether subjects thought using the same conditional probability concept that they had been presented with codes the transcripts from the memory test (bottom section of Table 25). In a repeated measures analysis of variance using only the unambiguous *inference* and *evidence* conditional probabilities, there was no significant difference in the number of mentions of p(e/h) versus p(h/e), nor a difference due to which conditional probability had been presented in the original problem. The interaction that tests the key hypothesis, that people recalled the conditional probability that they had read, was not significant ($F(1,26) = 1.32$, $p = .26$).

Comparison of the data from the fixed-order problem, the subject-selected order problem, and the memory test shows that the proportion of uses of the presented conditional probability expressions of reliability (Table 25) was similar on all three tasks[2]. Although this proportion is significantly different from random in the subject-selected order problem, presumably because subjects mentioned more concepts on this problem, still there is only 60% to 70% use of the conditional probability that is present on the page in

front of the subject, which represents only 20% to 40% better performance than random. This amount of error is strong support for the hypothesis that people confuse the conditional probabilities. The overall rate on the memory test is equivalent to that on the original problem, which suggests that "what you say is what you get", that is, there is no further slippage between the original verbalization of the concepts (excluding the actual reading) and the later recall.

3.9. Subjects' attention to the possibility of false positive evidence.

The coders' counts of the frequency with which subjects used each particular type of conditional probability concept (Table 22) are pertinent to the question whether subjects are alert to the possibility that the positive evidence may be false. On average subjects mentioned p(e/~h) 2.82 times in the entire session (3 problems plus memory test), which was more than the 1.02 times they mentioned p(e/h). Similarly, they mentioned p(~h/e) more often than p(h/e) (2.33 to 1.74 times, respectively). This suggests they recognize that it is important to consider false positives, when the idea has been presented to them. There was little discussion of either type of conditional probability when the reliability paragraph had not yet been presented (paragraph 3 of problems 1 and 3, the fixed-order problems).

4. Discussion.

The results of this study make it possible to select among the theories that explain people's performance on the diagnostic class of probabilistic inference word problems. The evidence favors the Confusion hypothesis. As we will see, this explains why subjects seem to neglect the base rate information in these problems, even though they may attend to it in other situations. It also gives a basis for predicting when people will follow the qualitative Bayesian principles. Analysis of the conditions of production and utilization of reliability information, in conjunction with what we know about the qualitative Bayesian principles, provides a framework for attempts to improve performance through decision aiding and training.

4.1. Evaluation of the three explanations of base rate neglect.

Our study of the three explanations for people's performance on probabilistic inference word problems, and for their neglect of base rate when the evidence counters it, has found little support for the theory that people neglect the base rate on principle or the theory that they explicitly interpolate between the probabilities associated with the base rate and the evidence. However, a number of analyses support the theory that people are confused by the conditional probability expression of the reliability of evidence.

4.1.1. Purposeful Neglect of Base Rate.

In its simplest form, this theory holds that people consider base rate information to be irrelevant and always ignore it. We found, however, that subjects used base rate as the answer when it alone was available. However, Cohen (1981, p 329) acknowledged that if the only available information is a statistical base rate, then it might be reasonable to attend to it. In this more complicated version of the theory, people still should not use the baserate information after evidence about the particular case has been received. But protocol analysis revealed that when subjects were presented with both evidence and base rate, one third of them used the base rate, either alone or in conjunction with the evidence information, and 42% of them used it when the reliability information was added. Others may have used it without saying so. It is not obvious how they could do this if they were purposefully neglecting base rate.

However, when reliability information was presented as the third piece of relevant information, a number of subjects used it as their answer. This is consistent with both the Cohen and Niiniluoto versions of this theory, as well as with the Confusion theory. Incidently, the fact that all four subjects who applied Bayes' Theorem did it using the base rate as p(h), instead of .50, is evidence against the Niiniluoto (1981) version of the theory, as is the finding that people are confused about the conditional probabilities.

In the theory, people will neglect a statistical base rate because they believe it is not pertinent to the particular case. They would be more justified in believing this about our Doctor and Insurance problems

than our Twins problem, where the base rate is stated in terms of these particular twins. If indeed people are sensitive to the scope of the given base rate, then when only base rate was available, more subjects should have used the base rate as their answer on the Twins problem than on the other two. But that did not happen. Subjects used the base rate as frequently on the Doctor problem as on the Twins problem, and only slightly less frequently on the Insurance problem (probably due to other features of that problem).

If base rate information is irrelevant to the issue, while evidence about the particular case is relevant, then subjects should recall the evidence more often than they recall the base rate. However, subjects recalled the two about equally frequently (and did not recall base rate more on the Twins problem than on the other two).

In sum, the data in this study offer little support for the theory that people purposefully ignore base rate information on probabilistic inference word problems.

4.1.2. Interpolation between base rate and evidence.

The second theory holds that people recognize the value of both the base rate and the unreliable evidence, and in order to produce their estimates of the probability of the hypothesis, they interpolate between the base rate and the 100% which is consonant with complete acceptance of the evidence. Tversky and Kahneman (1982) suggest that in this interpolation people underweight the base rate, and Bar-Hillel (1980) suggests that the weight on base rate depends on its perceived relevance. Another possible mechanism is that following the interpolation process, subjects may round to nearby probability landmarks or to numbers that are available in the word problem (Hamm, 1987c).

On the assumption that the process of interpolation would be explicit, the protocol analysis included categories for adjusting (from one anchor) and interpolation (between two anchors). However, such explicit strategies were seldom observed by the coders. In addition, subjects who used such strategies once seldom did so twice. It is possible that subjects were doing such processes implicitly. Almost all of the answers were "in range", between the low (base rate) and high (100%) anchors postulated for the interpolation, and so were consistent with an implicit interpolation process. However, this is not very diagnostic: most other processes would also produce answers in this range.

There was one class of integration strategy which could be clearly identified, and which some subjects used stably: Bayes' Theorem or its variants. Interestingly, most of these subjects did an incomplete variant, multiplying p(h) times p(e/h), which actually produces an answer that is outside the interpolation range. The subjects who used this strategy were a special class, mathematics graduate students who had been taught Bayes' Theorem, and so their stable use of integration strategies does not prove that the general population uses explicit interpolation processes.

In conclusion, we found no direct evidence that subjects produce their answers by explicitly integrating the base rate and the unreliable evidence. It is possible that they do an implicit integration process which is not identifiable in the transcripts, although other work shows that this is not a universally applied linear averaging (Hamm, 1987a; Ofir and Lynch, 1984).

4.1.3. Confusion of conditional probability expressions of reliability.

In contrast with the other two theories, a number of results support the Confusion theory. When the conditional probability was varied in the fixed-order problems, the conditional probability presented, p(e/h) or p(h/e), did not affect the number of subjects who used the conditional probability as their answer, nor the mean answer. The only subject who successfully applied the mechanics of Bayes' Theorem had actually received p(h/e), and so must have misrecognized it as p(e/h) before doing the calculation.

Process tracing analysis of the order in which people asked to be given the offered information, on the subject-selected order problem, showed that their preferences were inconsistent with their making a clear distinction between the two conditional probabilities. Instead, the subjects either needed to have them both before making an answer, considered the two to be identical, or were confused about the distinction between them.

Analysis of the subjects' thinking aloud also supported the Confusion theory. Five sixths of the subjects expressed difficulty understanding the difference between the two conditional probabilities on the subject-selected order problem. Two thirds of these ended up either concluding that they were identical, or failing to resolve the issue. Since the two concepts are not in fact identical, subjects who thought they were are also confused. Thus the protocol analysis data make it unlikely that subjects had strategies that needed both conditional probabilities.

When the subjects' verbalizations about reliability issues were coded into specific concepts, many of their statements were ambiguous with respect to which conditional probability was being expressed, and only 60% to 70% of those which could be identified were the same conditional probability p(e/h) or p(h/e) as was presented in the problem. This held in separate analyses of fixed-order and subject-selected order problems and of the memory test.

All this supports the theory's claim that people do not have a clear understanding of the difference between the two opposite conditional probabilities. The support is not quite as strong for the claim that subjects take p(e/h) to be p(h/e) and believe that p(h/e) is the appropriate response. While 75% of subjects used the conditional probability expression of reliability as their response on the Doctor problem, only 25% did on the Twins problem and none did on the Insurance problem. The theory's claim that subjects simply take p(e/h) to be p(h/e) and are otherwise fully confident that p(h/e) is an appropriate answer is not true on every probabilistic inference word problem. Perhaps their distrust of the source of the conditional probability information on some problems prevents them from using it in the hypothesized manner. The other possibility is a more general version of the Confusion hypothesis: subjects simply do not know what the conditional probabilities mean, nor what to do with them.

In conclusion, various results of this study converge in supporting the theory that the neglect of base rate, typical of performance on probabilistic inference word problems, is due to subjects not understanding the particular expression of the reliability of the evidence, p(e/h), nor how it should be applied to a situation (via Bayes' Theorem). Instead, many subjects mistakenly recognize it to be the general statement p(h/e) which they believe is directly pertinent to the question they have been asked.

4.2. Attention to general base rate or specific case information.

Psychology has demonstrated the overuse as well as the neglect of base rate (cf Jones, Worchel, Goethals, and Grumet, 1971). What governs whether people follow their prejudices and prior expectations, as opposed to being swayed by the particulars of the case? What is it about probabilistic inference word problems that makes subjects neglect prior expectations? The Confusion theory, so strongly supported by the results of this study, offers an answer to this question. It holds that in the particular set of situations that have been studied by the "diagnostic word problem" tradition, base rate is often neglected because when people are given information about the reliability of the case information (evidence) stated in conditional probability terms, they think this should be used as the answer because they do not know how to interpret the conditional probability correctly. That is, even though they appreciate the value of statistical base rate information, the conditional probability they are given seems to be, in itself, the correct answer to the question.

Under what conditions will this mechanism be engaged, and produce relative neglect of base rate? In the problems where it has been observed, subjects are forced to attend to explicit statements of the base rate, the evidence, and the unreliability of the evidence. (In contrast, in many situations the processing is more automatic. *Recognition* processes may give expectations (whether based on experience, or on ideology) greater weight in judgments than do *reasoning* processes. On the other hand, sometimes recognition is driven by similarity of features, as in the representativeness heuristic (Tversky and Kahneman, 1974; Dawes, 1986). The explicit probabilities in probabilistic inference word problems may make either of these processes unlikely.) These problems use an unfamiliar technical expression of reliability, the conditional probability. People may not know how to interpret its basic syntax: the probability of some event given some other event. And they may not know how to apply it in a formal inference procedure. Finally, the neglect of base rate and the use of the conditional probability as the response occur primarily when the evidence supports the unlikely hypothesis and the reliability of the evidence is

high (Ofir, 1988). Most real life uncertain inference situations do not possess all these special features, and so the confusion theory may not be pertinent to them. However, there are important contexts that have explicit probabilities, conditional probability expressions of reliability, and imperfectly reliable evidence for unlikely hypotheses (e.g., medicine: Eddy, 1982), and others where it has justifiably been advocated that conditional probability measures of evidence reliability should be made available (e.g., legal testimony; Koshland, 1988, and Schum, 1987; but see Cohen, 1977, and Tribe, 1971). For these situations, the Confusion theory is important.

4.2.1. The production and utilization of the p(e/h), p(h), and p(h/e) statistics.

People whose professions demand they use statistics in probabilistic inference may use them in the same way that our subjects have done. Consideration of the production and utilization of statistics in medicine and clinical psychology, for example, shows that conditions are just right to produce the same confusion between conditional probabilities that has been observed in this study (Eddy, 1982; Dawes, 1986; Widiger, et al, 1984).

Statistics about base rate p(h) and about the reliability of tests, p(e/h) and p(h/e), are produced in scientific studies. Tests are calibrated in laboratory studies, where it is equally easy to produce p(e/h) and p(h/e) statements. Base rates are produced in epidemiological studies, which require less scientific control and can be done frequently to keep up with changing conditions and population differences. Only if everyone in a population has been given the test e can epidemiological studies produce p(h/e) statistics. P(h) and p(h/e) are statistics of limited generality. They are true only in situations that have the same base rate as the sample in the scientific study that produced them. On the other hand, p(e/h) is much more applicable as a measure of the reliability of a test that produces evidence e (see Meyer and Pauker, 1987).

Because most scientists (aided by peer review) understand that only p(e/h) is a generally applicable measure of reliability, they publish and disseminate reliability information in the p(e/h) form. Producing p(h/e) statistics in the laboratory requires as much work as producing p(e/h), but these are of only limited generality, so they are seldom produced and published (except by mistake; see Eddy, 1982). Statistics on base rate p(h) are updated on an ongoing basis by epidemiological studies.

In the clinical context where this knowledge is applied, it is p(h/e) (for that application situation) that is needed. The sophisticated knowledge user knows that p(e/h) is generally applicable. Its application requires that it be integrated with information about p(h), using Bayes' Theorem. In contrast with this norm, many people who need to know p(h/e) for a case misrecognize the p(e/h) that is available in the literature as p(h/e) (Eddy, 1982; Dawes, 1986), just as did the subjects in this study.

One factor contributing to such mistakes in the application of medical statistics is that formal medical education is focused on "basic science", while the education in the applied science of the clinic is informal (Hammond and Hamm, 1983). Therefore, most medical students are formally trained to think about the laboratory production of medical statistics (where p(e/h) and p(h/e) are equally of interest), but not about the application of those medical statistics [where p(h/e) must be derived by working with p(e/h) and p(h)]. This could be corrected through training (see texts such as Weinstein et al, 1980) and through the design of systems for producing and utilizing reliability information.

4.2.2. Influence of formal training on people's vulnerability to conditional probability confusion.

Probability is an abstract symbolic language. Although it has entered everyday discourse in a number of realms (e.g., sports, weather), the interpretation of the conditional probabilities presented in these word problems may require special education. We compared undergraduates, presumably untrained in probabilistic inference, with mathematics graduate students, who have studied the formal Bayes' Theorem technique. Only 4 of the 14 mathematics graduate students mentioned the applicability of Bayes' Theorem to the problems. None of them applied it perfectly. Three of them multiplied p(h) by p(e/h), but neglected the possibility of a false alarm, p(e/~h)*p(~h), and so missed the point of the ratio in Bayes' Theorem. Thus despite their training they violated the 4th qualitative Bayesian principle (to attend to false positive), although they obeyed the 3rd (to attend to reliability). The fourth subject who attempted Bayes' Theorem mistakenly interpreted the presented p(h/e) conditional probability as p(e/h) in an

otherwise correct application of the formula.

Besides their attempts to apply Bayes' Theorem, the mathematics graduate students had statistically more sophisticated intuitions about the problems, as would be expected (see Nisbett, Krantz, Jepson, and Kunda, 1983). They tended to ask for the base rate first [Principle 2] on the subject-selected order problem, while undergraduates asked for evidence first [Principle 1]. On the fixed-order problems they tended to adopt the cognitively complex strategy of using base rate in conjunction with other information, while undergraduates used it alone or ignored it completely. Thus the graduate students were more likely to obey the 5th qualitative Bayesian principle: that evidence and base rate should be integrated.

4.2.3. The possibility of stable individual differences.

Data in this study, as in all probabilistic inference word problem studies, are quite variable. Answers come from the full 0 to 1 range of probabilities. Although we occasionally compared the means of groups of subjects, we more frequently counted subjects whose answers were consistent with the use of various strategies and used X^2 statistics to see whether one occurred more frequently. But although significantly more subjects may use one strategy than another, this does not mean that alternative strategies were never used. There is the possibility that different subjects have different stable strategies. However, except for those mathematics graduate students who used variants of Bayes' Theorem, there was little evidence of stable strategy use here.

4.2.4. Differences between problems.

Although subjects were equally confused about the conditional probabilities on the Doctor, Insurance, and Twins problems, there was wide variation in the use of the conditional probability as the response. It happened frequently (75%) with the Doctor problem, infrequently (25%) with the Twins problem, and never with the Insurance problem. It may depend on the authority of the reliability information (a medical textbook, a mother's testimony, and the result of psychological research, respectively), as well as on the perceived relevance of the base rate information (Bar-Hillel, 1980). Since we have not systematically explored variations in problem content we can offer no general conclusion on this topic, but it remains an important factor in interpretation of all work using probabilistic inference word problems.

4.3. Implications.

We conclude by considering the implications of our findings concerning whether people behave in accord with the qualititative Bayesian principles. We have shown that people find it natural to attend to the given evidence (Principle 1) on these sorts of word problems, despite the prejudice and stereotypy that appear in other contexts. We have confirmed the findings (Ofir, 1988, and others) that people appreciate the pertinence of base rate information (Principle 2) in most cases. This contradicts the argument of Cohen (1981). We have shown that the apparent exception, the neglect of base rate (Bar-Hillel, 1980; Tversky and Kahneman, 1982), is due to people's difficulties in interpreting conditional probability expressions of reliability, rather than to a lack of appreciation of the base rate. Subjects recognized the pertinence of the reliability of the evidence (Principle 3), although they did not know how to use probabilistic measures of reliability. They spoke about the possibility of a false alarm p(e/~h) more frequently than about p(e/h) when both ideas were presented to them, which shows that they recognize the pertinence of Principle 4. However, 3 of 4 graduate students who tried to use Bayes' Theorem left out the part that deals with false alarms, and previous work (Doherty, Mynatt, Tweney, and Schiavo, 1979) suggests they don't consider false alarms unless prompted. Subjects often did not integrate information about the base rate and unreliable evidence (Principle 5). They occasionally used numbers that represent attending to the base rate alone, or to the evidence alone. Even when their responses were in between the base rate and 100%, this often represented the use of a conditional probability p(e/h) that they misrecognized as p(h/e), rather than a subjective integration of the competing types of information. Finally, Principle 6 holds that people should make reasonable assumptions concerning information that is missing. Although we did not address this directly in the study, subjects seemed to make simplifying assumptions, rather than making their problems more complex by considering factors that have not been mentioned and

assigning reasonable values to them. For example, on only a third of the problems did the subjects spontaneously consider the reliability of the evidence, before the specific reliability information was given to them.

In summary, people tend to follow the early principles on the list better than the later ones. Their failures may be traced to (a) not knowing the principle or not knowing a procedure for carrying it out, (b) not being reminded of the principle, or (c) misinterpreting information (conditional probabilities) and deciding that a principle (Principle 2) is no longer applicable given that misinterpretation. This view has implications for improving performance through training and decision aiding, and for evaluating the performance of any human or man-machine inferencing system.

Training. It should be possible to conduct training in probabilistic inference by building on the basic appreciation that people have for evidence, base rate, and reliability information. Emphasis should be placed on three areas:

1. People should be made alert to the possibility of false positive evidence, which they often neglect.
2. People should be taught how to integrate prior expectations and current evidence, either through Bayes' Theorem or through appropriate estimation techniques that are responsive to the reliability of the evidence.
3. People should be taught to correctly interpret and use reliability information, including avoiding the errors of
   a. misconceiving a p(e/h) probability as a p(h/e), and
   b. assuming that a p(h/e) produced elsewhere is applicable to the present situation.

   In this way they will not be induced to ignore base rate information.

Aids such as the 2 by 2 table explored by Lichtenstein and MacGregor (1984) can sharpen the distinction between the conditional probabilities and also remind people of the possibility of false positive evidence.

Development of expertise. The long term goal of fostering expertise can be distinguished from the short term goal of training someone to perform a particular decision making task. Dreyfus and Dreyfus (1986; see Hamm, 1988b) describe how expertise is developed through repeated engagements with a task in which an analytic perspective is provided by teachers (Bayes' Theorem in the case of probabilistic inference). They warn against the illusion that intuitive expert performance can be accurate without this sort of long term analytical engagement.

Decision aiding. Decision aids should be designed so that they do not give users the opportunity to misinterpret conditional probabilities and thus enter wrong information into the system. Otherwise, the basic decision aiding approach seems well founded: to remind people about prior expectations, the possibility of false alarms, and the unreliability of evidence (or to automatically take these into account), and to reliably apply the probability calculus if the needed information is available. There is a need for aids in situations where complete information about the probabilistic structure of the environment is not available. The qualitative Bayesian principles may be useful for building such aids. However, decision aids also need to be sensitive to variations in the universality of p(e/h) and p(h) statistics. In addition, non-Bayesian techniques such as Collins and Michalski's (1987) plausible inference or Fox, O'Neil, Glowinski, and Clark's (1988) symbolic inference schemata are being developed in an attempt to provide alternative ways to make inferences using data bases that lack well-measured relative frequencies and conditional probabilities.

Evaluation. It is increasingly becoming necessary to evaluate the inference processes of individuals, groups, or man-machine systems. The results of this study suggest that it might be useful not only to ask whether inference is consistent with Bayes' Theorem or at least with the qualitative Bayesian principles, but also whether appropriate distinctions between the different classes of conditional probability are being made.

References.

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

Beyth-Marom, R., and Fischhoff, B. (1983). Diagnosticity and pseudodiagnosticity. J. of Personality and Social Psychology, 45, 1185-1195.

Birnbaum, M.H., and Mellers, B.A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. J. Personality and Social Psychology, 45, 792-804.

Cohen, J. (1980). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cohen, L.J. (1977). The Probable and the Provable. Oxford: Oxford University Press.

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

Cohen, M., Schum, D.A., Freeling, A.N.S., and Chinnis, J.O., Jr. (1985). On the art and science of hedging a conclusion: Alternative theories of uncertainty in intelligence analysis (Technical Report 84-6). Reston, VA: Decision Science Consortium.

Collins, A., and Michalski, R. (1987). The logic of plausible reasoning: A core theory. Cambridge, MA: Bolt, Baranek, and Newman, Inc.

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

Dellarosa, D., and Bourne, L.E., Jr. (1984). Decision and memory: Differential retrievability of consistent and contradictory evidence. J. Verbal Learning and Verbal Behavior, 23, 669-682.

Doherty, M.E., Mynatt, C.R., Tweney, R.D., and Schiavo, M.D. (1979). Pseudodiagnosticity. Acta Psychologica, 43, 111-121.

Dreyfus, H.L., and Dreyfus, S.E. (1986). Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer. New York: The Free Press.

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

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), Formal representation of human judgment. New York: Wiley, pp 17-52.

Edwards, W., Lindman, H., and Savage, L.J. (1963). Bayesian statistical inference for psychological research. Psychological Review, 70, 193-242.

Ericsson, K.A., and Simon, H.A. (1984). Protocol analysis: Verbal reports as data. Cambridge, Mass.: The MIT Press.

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

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

Fox, J., O'Neil, M., Glowinski, A.J., and Clark, D. (1988). A logic of decision making. Illinois Interdisciplinary Workshop on Decision Making, Champaign-Urbana, Illinois, June.

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

Hamm, R.M. (1987b). Explanations of the use of reliability information as the response in probabilistic inference word problems. (Publication #87-13.) Institute of Cognitive Science, University of Colorado, Boulder.

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

Hamm, R.M. (1988a). Accuracy of probabilistic inference using verbal *vs* numerical probabilities. Psychonomics Society meetings, Chicago.

Hamm, R.M. (1988b). Clinical intuition and clinical analysis: Expertise and the cognitive continuum. In Jack Dowie and Arthur Elstein (Eds.), Professional Judgment: A Reader in Clinical Decision Making. Cambridge: Cambridge University Press, pp 78-105.

Hamm, R.M., Lusk, C.M., Miller, M.A., Smith, D.L., and Young, I.E. (1988). Coder's materials and reliabilities for analysis of thinking aloud protocols from study on the use of conditional probabilities in probabilistic inference. (Publication #88-13.) Institute of Cognitive Science, University of Colorado, Boulder.

Hammond, K.R., and Hamm, R.M. (1983). Thoughts on the acquisition and application of medical knowledge. In C.P. Friedman and E.F. Purcell, (Eds.), The New Biology and Medical Education: Merging the Biological, Information, and Cognitive Sciences. New York: Josiah Macy, Jr., Foundation, pp. 190-197.

Jones, E.E., Worchel, S., Goethals, G.R., and Grumet, J.F. (1971). Prior expectancy and behavioral extremity as determinants of attitude attribution. J. Experimental Social Psychology, 7, 59-80.

Kahneman, D., and Tversky, A. (1972). On prediction and judgment. Oregon Research Institute Research Monograph, 12(4).

Koshland, D. (1988). A tale of two techniques (editorial). Science, 242, 993.

Kozminsky, E., Kintsch, W., and Bourne, L.E., Jr. (1981). Decision making with texts: Information analysis and schema acquisition. J. Experimental Psychology: General, 110, 363-380.

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

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

McClelland, G.H., Stewart, B.E., Judd, C.M., and Bourne, L.E., Jr. (1987). Effects of choice task on attribute memory. Organizational Behavior and Human Decision Processes, 40, 235-254.

Medin, D.L., and Edelson, S.M. (1988). Problem structure and the use of base-rate information from experience. J. Experimental Psychology: General, 117, 68-85.

Meyer, K.B., and Pauker, S.G. (1987). Screening for HIV: Can we afford the false positive rate? New England Journal of Medicine, 317, 238-241.

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

Nisbett, R.E., Krantz, D.H., Jepson, C., and Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. Psychological Review, 90, 339-363.

Ofir, C. (1988). Psuedodiagnosticity in judgment under uncertainty. Organizational Behavior and Human Decision Performance, 42, 343-363.

Ofir, C., and Lynch, J.G., Jr. (1984). Context effects on judgment under uncertainty. J. Consumer Research, 11, 668-679.

Payne, J. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. Organizational Behavior and Human Performance, 16, 366-387.

Pollatsek, A., Well, A.D., Konold, C., Hardiman, P., and Cobb, G. (1987). Understanding conditional probabilities. Organizational Behavior and Human Decision Processes, 40, 255-269.

Rowe, H.A.H. (1985). Problem Solving and Intelligence. Hillsdale, N.J.: Erlbaum.

Scott, W.A. (1955). Reliability of content analysis: The case of nominal scaling coding. Public Opinion Quarterly, 19, 321-325.

Schum, D.A. (1987). Evidence and Inference for the Intelligence Analyst. Lanham, MD: University Press of America.

Slovic, P., and Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organizational Behavior and Human Performance, 6, 649-744.

Tribe, L. H. (1971). Trial by mathematics: Precision and ritual in the legal process. Harvard Law Review, 84, 1329-1393.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.

Tversky, A., and Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), Progress in Social Psychology. Hillsdale, N.J.: Lawrence Erlbaum Assoc., Inc. [Reprinted in D. Kahneman, P. Slovic, and A. Tversky (Eds.), Judgment under Uncertainty: Heuristics and Biases. New York: Cambridge University Press, 1982, pp 117-128.]

Tversky, A., and Kahneman, D. (1982). Evidential impact on 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.

von Winterfeldt, D., and Edwards, W. (1986). Decision analysis and behavioral research. New York: Cambridge University Press.

Weinstein, M.C., Fineberg, H.V., Elstein, A.S., Frazier, H.S., Neuhauser, D., Neutra, R.R., and McNeil, B.J. (1980). Clinical Decision Analysis. Philadelphia: W.B. Saunders Company.

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.

Zwick, R. (1988). Another look at interrater agreement. Psychological Bulletin, 103, 374-378.

Tables.

Table 1.
Coding Categories for Fixed-Order Problems.

1a. mentions correct base rate number
1b. mentions incorrect base rate number
1c. mentions nonspecific base rate
1d. no mention of base rate

2a. uses base rate as the answer
2b. uses base rate in an adjustment process (list adjustment process)
2c. no use of base rate (list reason)

3a. interpolation - no justification
3b. interpolation - evidence justification
3c. interpolation - reliability justification
3d. interpolation - other justification (list justification)
3e. anchor and adjust - no justification
3f. anchor and adjust - evidence justification
3g. anchor and adjust - reliability justification
3h. anchor and adjust - other justification (list justification)
3i. some other adjustment is made (list adjustment process)
3j. no adjustment is made

4a. uses base rate only - no justification
4b. uses base rate only - reliability justification
4c. uses base rate only - other justification (list justification)
4d. no use of base rate (list reason)
4e. uses base rate in conjunction with other info (circle something in part 3)

5a. mentions correct evidence
5b. mentions incorrect evidence
5c. no mention of evidence

6a. uses evidence only
6b. uses evidence to adjust (circle something in part 3)
6c. no use of evidence - reliability justification
6d. no use of evidence (list justification)

7a. mentions reliability issue - goes beyond provided info
7b. mentions reliability issue - repeats info provided
7c. no mention of reliability

8a. mentions correct reliability number
8b. mentions incorrect reliability number
8c. mentions nonspecific reliability
8d. no mention of reliability

9a. uses "correct" reliability only
9b. uses "error" reliability only
9c. uses reliability to adjust (circle something in part 3)
9d. no use of reliability (list)

10a. uses Bayes' theorem
10b. mentions Bayes' theorem but doesn't use
10c. no use of Bayes' theorem

Table 2
Coding sheet for conditional probability coding schemes.

	Problem 1		Problem 2				Problem 3		Memory
			Ordering		Answering				
	Par 3	Par 4	Par 3	Par 4	Par 3	Par 4	Par 3	Par 4	Test
p(h/e).									
p(h/~e).									
p(~h/e).									
p(~h/~e).									
p(e/h).									
p(e/~h).									
p(~e/h).									
p(~e/~h).									
p(correct evidence)									
p(wrong evidence)									
p(correct inference)									
p(wrong inference)									
p(h & e)									
p(h & ~e)									
p(~h & e)									
p(~h & ~e)									
Ambiguous.									
No mention.									

Note: Instructions were: "At each location for which reliability interpretation coding and counting is required, identify each statement of the reliability concepts as one of the following interpretations. If more than one interpretation is stated, mark each. If an interpretation is stated more than once, put a mark for each time, so that we have a count of how many times they used the idea."

Table 3.
The specific concepts called "e" and "h" in each problem.

	Twins	Doctor	Insurance
h	was Stephen	is toxic uremia	have good past driving record
~h	was Paul	is hepatitis	have bad past driving record
e	say Stephen	test says t.u.	says he has a good past driving record
~e	say Paul	test says hep.	admit to a bad past driving record

Table 4.
Combinations of coding categories into broader categories.

Evidence/Inference Consistent/Incons. Specific	Evidence/Inference Consistent/Incons. General	Evidence/Inference General
p(h/e)		
p(~h/~e)	p(correct inference)	
p(correct inference)		p(inference)
p(h/~e)		
p(~h/e)	p(wrong inference)	
p(wrong inference)		
p(e/h)		
p(~e/~h)	p(correct evidence)	
p(correct evidence)		p(evidence)
p(e/~h)		
p(~e/h)	p(wrong evidence)	
p(wrong evidence)		
p(h and e)		
p(~h and ~e)	p(consistent conj)	
p(h and ~e)		p(ambiguous)
p(~h and e)	p(inconsist. conj)	
Ambigous	Ambigous	
No Conditional Prob[a]	No Conditional Prob	No Conditional Prob

[a]Either the sentence was not considered to express a probability involving evidence and hypothesis, or its referent paragraph could not be determined.

Table 5.
Comparison of coders' categorization judgments to determine reliability of conditional probability categorization scheme.

Coder 3's Coding

	Evidence						Inference				Conjunction			
	The Probability of						The Probability of				The Probability of			
Coder 4's Coding	e/h	e/~h	~e/h	~e/~h	ce	we	h/e	h/~e	~h/e	ci	h&e	h&~e	~h&e	Total
Evidence														
P(e/h)	1	0	0	0	0	0	1	0	0	0	2	0	0	4
P(e/~h)	0	16	1	0	1	1	0	0	5	0	0	0	0	24
P(~e/h)	0	0	8	0	0	0	0	1	0	0	0	0	0	9
P(~e/~h)	0	0	0	0	0	0	0	0	0	0	0	0	1	1
P(correct evidence)	0	0	0	0	4	0	1	0	0	0	0	0	0	5
P(wrong evidence)	0	1	0	0	0	2	1	0	0	0	0	0	0	4
Inference														
P(h/e)	1	0	0	0	1	0	6	0	2	0	1	0	0	11
P(h/~e)	0	0	0	0	0	0	0	0	0	0	0	1	0	1
P(~h/e)	0	1	0	0	0	0	0	0	10	0	0	0	0	11
P(correct inference)	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Ambiguous														
P(h&e)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P(h&~e)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P(~h&e)	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	2	18	9	0	7	3	9	1	17	0	3	1	1	71

Table 6
Mean within subject correlations between coders, within category, over locations.

Most specific categories, as originally coded.

	Mn r	SD r	Min	Max	N
p(h/e)	.702	.321	.15	1.0	6
p(h/~e)	1.000	.	1.00	1.0	1
p(~h/e)	.637	.485	-.27	1.0	7
p(~h/~e)	.	.	.	.	0
p(e/h)	.884	.	.88	.9	1
p(e/~h)	.527	.472	-.19	1.0	8
p(~e/h)	.719	.563	-.13	1.0	4
p(~e/~h)	.	.	.	.	0
p(correct evidence)	.112	.436	-.29	.7	5
p(wrong evidence)	.591	.448	.11	1.0	3
p(correct inference)	.	.	.	.	0
p(wrong inference)	.	.	.	.	0
p(h&e)	1.0	.	1.00	1.0	1
p(h&~e)	.	.	.	.	0
p(~h&e)	.	.	.	.	0
p(~h&~e)	.	.	.	.	0
ambiguous	.128	.381	-.61	.7	10
no mention	.675	.214	.40	1.0	6

General categories, inference/evidence and consistent/inconsistent

correct evid	.076	.397	-.26	.66	6
wrong evid	.534	.446	-.22	1.00	9
correct inf	.560	.475	-.29	1.00	7
wrong inf	.761	.466	-.27	1.00	7
correct conj	1.000	.	1.00	1.00	1
wrong conj	~	~	~	~	~
"ambig"	.128	.381	-.61	.70	10
no mention	.675	.214	.40	1.00	6

General categories, inference/evidence

evidence	.416	.457	-.24	.97	9
inference	.624	.454	-.29	1.00	10
ambiguous	.159	.394	-.55	.70	10
no mention	.675	.214	.40	1.00	6

Table 7.
Number of subjects who used the base rate information, after each paragraph of information, for fixed-order problems.

	Problem		
Information Given[a]	Doctor	Insurance	Twins
Problem definition	0	0	0
Base rate	16	9	16
Base rate & Evidence	0	3	3
Base rate, Evidence & Conditional Probability	2	4	1
	N=20	N=21	N=24

[a] The base rate information was not available in the first row.

Table 8.
Number of subjects who completely accepted the evidence information and used a response of 100%, after each paragraph of information, on fixed-order problems.

	Problem		
	Doctor	Insurance	Twins
Information Given[a]			
Problem definition	0	2	0
Base rate	0	0	0
Base rate & Evidence	10	4	3
Base rate, Evidence & Conditional Probability	0	0	1
	N=20	N=21	N=24

[a] The evidence information was available in only the last two rows.

Table 9.
Number of subjects who used the conditional probability information as their response, when base rate, evidence, and conditional probability information had been given, on fixed-order problems.

	Use of Conditional Probability		
	Correct[a]	Complementary[b]	
Doctor Problem, N=20			
P(e/h) presented	8	0	N=11
P(h/e) presented	7	0	N=9
Insurance Problem, N=21			
P(e/h) presented	0	1	N=13
P(h/e) presented	0	0	N=8
Twins problem, N=24			
P(e/h) presented	3	0	N=9
P(h/e) presented	3	2	N=15

[a]The conditional probability stated in the problem to hold between the evidence and the hypothesis: p(e/h) or p(h/e).

[b]The complement of the stated probability: p(~e/h) or p(~h/e).

Table 10
Number of subjects choosing to receive information in each order, on subject-selected order problem.

Information Order	Frequency	Percent
b e peh phe	3	9.4
b e phe peh	6	18.8
b peh phe e	3	9.4
b phe peh e	3	9.4
e b peh phe	2	6.3
e b phe peh	7	21.9
e peh b phe	1	3.1
e peh phe b	3	9.4
e phe b peh	1	3.1
e phe peh b	2	6.3
phe peh e b	1	3.1
Total	32	100.0

Note: b = base rate, e = evidence, peh = p(evidence/hypothesis), and phe = p(hypothesis/evidence).

Table 11
The ordinal position in which base rate and evidence were selected, for the three problems.

		Position Selected			
	N	1	2	3	4
Base rate					
Doctor	12	3	4	1	4
Insurance	11	6	5	0	0
Twins	9	6	0	1	2
Total	32	15	9	2	6
Evidence					
Doctor	12	9	3	0	0
Insurance	11	5	2	4	0
Twins	9	2	4	1	2
Total	32	16	9	5	2

Note: b = base rate, e = evidence, peh = p(evidence/hypothesis), and phe = p(hypothesis/evidence).

Table 12
The ordinal position in which base rate and evidence were selected, for the different subject types.

		Position Selected			
	N	1	2	3	4
Base rate					
Undergraduate	15	6	5	1	3
Grad. Student	14	9	2	1	2
Professional	3	0	2	0	1
Total	32	15	9	2	6
Evidence					
Undergraduate	15	8	3	1	3
Grad. Student	14	5	6	0	3
Professional	3	3	0	0	0
Total	32	16	9	1	6

Table 13
Relation between subjects' distinctions between the conditional probabilities, and their strategies for acquiring information.

Distinction	Constraints on orders	Info gained from order S chose	Influence of arbitrary factor on order
Purposeful	None	Relative importance	None
Both needed	Adjacent	None	High
Identical	One at end	None	High
Confused	Adjacent	None	High

Table 14
Those types of distinction between conditional probabilities with which each information preference order is compatible.

Preference order	Number of S with each pref order	Cond Prob Relations	One last non-adjac	Adjacent	Adjacent Not last	p(e/h) first	p(h/e) first
		Compatible Types of Distinction	Identical	Identical Confused or B. N.	Confused or Both Needed	Purposeful	
b e peh phe	3		0	3	0	3	~
b e phe peh	6		0	6	0	~	6
b peh phe e	3		0	0	3	3	~
b phe peh e	3		0	0	3	~	3
e b peh phe	2		0	2	0	2	~
e b phe peh	7		0	7	0	~	7
e peh b phe	1		1	0	0	1	~
e peh phe b	3		0	0	3	3	~
e phe b peh	1		1	0	0	~	1
e phe peh b	2		0	0	2	~	2
phe peh e b	1		0	0	1	~	1
Totals	32		2	18	12	12	20
# expected by chance	~		10.7	5.3	10.7	16	16

Table 15
Relation between order in which conditional probabilities are listed and order in which they are requested.

		Paragraph selected first		
		p(e/h)	p(h/e)	
Paragraph listed first	p(e/h)	8	6	14
	p(h/e)	4	14	18
		12	20	32

Table 16
Mentioning and use of base rate at paragraphs 2, 3, and 4 of the fixed-order problems.

Categorization of base rate use after base rate paragraph (#2).

	Doctor N=20	Insurance N=19	Twins N=20
Mentions Correct Base Rate	19	18	18
Mentions Incorrect Base Rate	0	0	0
Mentions Nonspecific Base Rate	1	0	2
No Mention of Base Rate	0	1	0
Uses Base Rate as Answer	16	9	15
Uses Base Rate to Adjust Answer	1	5	3
No use of Base Rate	3	3	1
Not Categorized in this Scheme	0	2	1

Categorization of base rate use after evidence paragraph (#3).

	Doctor N=20	Insurance N=19	Twins N=20
Mentions Correct Base Rate	1	8	8
Mentions Incorrect Base Rate	0	1	0
Mentions Nonspecific Base Rate	1	0	2
No Mention of Base Rate	18	10	10
Uses Base Rate Only- No Justification	0	1	2
Uses Base Rate Only- Reliability Justification	0	1	2
Uses Base Rate Only- Other Justification	0	2	1
Uses Base Rate in Conjunction With Other Information	1	4	4
No use of Base Rate	17	10	10
Not Categorized in this Scheme	2	1	1

Categorization of base rate use after conditional probability paragraph (#4).

	Doctor N=20	Insurance N=19	Twins N=20
Mentions Correct Base Rate	6	8	11
Mentions Incorrect Base Rate	0	1	0
Mentions Nonspecific Base Rate	0	1	4
No Mention of Base Rate	13	9	5
Uses Base Rate Only- No Justification	0	1	0
Uses Base Rate Only- Reliability Justification	0	2	2
Uses Base Rate Only- Other Justification	0	1	0
Uses Base Rate in Conjunction With Other Information	3	4	10
No use of Base Rate	16	10	7
Not Categorized in this Scheme	2	1	1

Note: each subject did two problems in this table. For each problem, they answered after each of the three paragraphs. For each paragraph, their transcripts were coded in two separate categorization schemes: mentioning, and use, of base rate.

Table 17
Mean answer for subjects who did or did not mention base rate, after paragraphs 2, 3, and 4 for fixed-order problems.

	Problem Type								
	Doctor			Insurance			Twins		
N	20			19			20		
	Ans[a]	Acc[b]	Use[c]	Ans	Acc	Use	Ans	Acc	Use
Base rate paragraph									
Mentions	.18	.04 n = 20	16	.39	.11 n = 18	8	.32	.12 n = 20	13
Doesn't Mention	~	~ n = 0	~	.8000	.4500 n = 1	0	~	~ n = 0	~
Evidence paragraph									
Mentions	.88	~[d] n = 2	0	.47	~ n = 9	1	.69	~ n = 10	2
Doesn't Mention	.97	~ n = 18	0	.79	~ n = 10	2	.76	~ n = 10	0
Reliability paragraph									
Mentions	.52	.35 n = 6[a]	2	.37	.32 n = 10	2	.46	.24 n = 15	1
Doesn't Mention	.81	.16 n = 13	0	.56	.24 n = 9	2	.62	.20 n = 5	0

[a]Tape recorder ran out before reliability paragraph for one subject. [a]Ans: the mean answer of all subjects in this category.

[b]Acc: the mean accuracy (absolute deviation of subject's answer from the correct answer, which depends on the information available and on whether the reliability is p(e/h) or p(h/e), and which is indeterminate when only base rate and evidence have been given).

[c]Use: the number of subjects in this category who used the base rate as their response.

[d]There is no unique correct answer when only base rate and evidence have been presented.

Table 18
Mentioning and use of evidence, paragraphs 3 and 4, fixed-order problems.

Categorization of evidence use after evidence paragraph (#3).

	Doctor N=20	Insurance N=19	Twins N=20
Mentions Correct Evidence	19	17	19
No Mention of Evidence	1	2	1
Uses Evidence Only as Answer	16	1	3
Uses Evidence to Adjust Answer	3	5	7
No Use of Evidence- Reliability Justification	0	3	6
No use of Evidence- Other or No Justification	1	10	4

Categorization of evidence use after conditional probability paragraph (#4).

	Doctor N=20	Insurance N=19	Twins N=20
Mentions Correct Evidence	10	8	16
No Mention of Evidence	10	11	4
Uses Evidence Only as Answer	0	0	2
Uses Evidence to Adjust Answer	3	2	6
No Use of Evidence- Reliability Justification	12	6	5
No use of Evidence- Other or No Justification	5	11	7

Note: each subject appears in this table on two different problems, two paragraphs each, two categorization schemes per paragraph.

Table 19
Mentioning and use of reliability concept, paragraphs 3 and 4, fixed-order problems.

Categorization of use of reliability concept after evidence paragraph (#3).

	Doctor N=20	Insurance N=19	Twins N=20
Mentions Reliability Issue -- Goes Beyond Provided Information	11	3	4
Mentions Reliability Issue -- Repeats Information Provided	0	3	1
No Mention of Reliability	9	13	15

Categorization of use of reliability concept after conditional probability paragraph (#4).

	Doctor		Insurance		Twins	
	p(e/h) N=11	p(h/e) N=9	p(e/h) N=10	p(h/e) N=9	p(e/h) N=8	p(h/e) N=12
Mentions Correct Reliability Number	11	9	8	8	4	9
Mentions Incorrect Reliability Number	0	0	0	0	0	1
Mentions Reliability Without Specific Number	0	0	1	0	1	0
No Mention of Reliability	0	0	1	1	3	2
Uses Correct Reliability Number	8	7	0	0	0	2
Uses Complement of Reliability Number	0	0	0	0	0	1
Uses Reliability to Adjust Answer	3	1	6	5	5	4
No Use of Reliability	0	1	4	4	3	5

Note: in this table each subject did two problems, one with p(e/h) and one with p(h/e), and answered on two paragraphs per problem, and was coded on one (paragraph 3) or two (paragraph 4) coding schemes.

Table 20
Differences between subject types' use of conditional probability information on paragraph 4 of fixed-order problems.

	Undergraduate	Graduate	Professional
Reliability used as answer	11	6	1
Reliability used in adjustment	9	15	0
No use of reliability	9	3	5

Table 21
Number of subjects who recalled each concept after each of four questions in the memory test.

Question	Base rate	Evid-ence	Reliability p(e/h)	Reliability p(h/e)	Final Answer
Openended					
no mention	17	17	9	6	22
mentions	3	6	0	1	0
incorrect #	0	~	0	0	1
correct #	3	~	2	5	0
N =	23	23	11	12	23
Key information?					
no mention	8	13	8	4	26
mentions	10	15	2	4	2
incorrect #	3	~	2	0	0
correct #	7	~	1	7	0
N =	28	28	13	15	28
Final answer?					
no mention	26	25	11	14	1
mentions	0	3	0	0	2
incorrect #	0	~	0	0	9
correct #	2	~	2	1	16
N =	28	28	13	15	28
How got answer?					
no mention	15	14	8	5	16
mentions	6	14	3	4	4
incorrect #	1	~	1	0	3
correct #	6	~	1	6	5
N =	28	28	13	15	28

Note: each subject appeared once in each column of each subtable, except each appeared in only one of the two reliability columns.

Table 22
Average number of times each subject used each category of conditional probability, at each location in the transcript.

Original Categorization: Evidence/Inference, Consistent/Inconsistent, Specific Hypotheses.

	Problem 1		Problem 2				Problem 3		Memory
	Par 3	Par 4	Ordering		Answering		Par 3	Par 4	Test
			Par 3	Par 4	Par 3	Par 4			
p(h/e).	.000	.267	.033	.033	.400	.276	.034	.483	.214
p(h/~e).	.000	.000	.033	.033	.100	.034	.000	.103	.000
p(~h/e).	.000	.600	.133	.033	.800	.345	.000	.379	.036
p(~h/~e).	.000	.000	.000	.000	.000	.034	.000	.000	.000
p(e/h).	.000	.233	.033	.067	.200	.276	.000	.138	.071
p(e/~h).	.000	.433	.167	.267	.500	.655	.138	.552	.107
p(~e/h).	.000	.033	.033	.033	.367	.069	.000	.069	.000
p(~e/~h).	.000	.067	.000	.000	.033	.000	.000	.000	.000
p(correct evidence)	.033	.200	.300	.133	.067	.034	.034	.138	.357
p(wrong evidence)	.033	.100	.200	.033	.000	.034	.069	.103	.071
p(correct inference)	.000	.000	.000	.033	.033	.103	.000	.034	.143
p(wrong inference)	.000	.000	.000	.000	.000	.000	.000	.034	.036
p(h & e)	.000	.033	.000	.000	.067	.034	.000	.103	.000
p(h & ~e)	.000	.000	.000	.000	.000	.000	.000	.000	.000
p(~h & e)	.000	.000	.033	.033	.033	.000	.000	.000	.000
p(~h & ~e)	.000	.000	.000	.000	.000	.000	.000	.000	.000
Ambiguous	.700	2.033	.900	.733	1.400	1.310	.655	1.379	1.107
No mention	.467	.000	.000	.000	.000	.000	.310	.034	.179
N	30	30	30	30	30	29	29	29	28

Combined Categories: Evidence/Inference, Consistent/Inconsistent.

	Problem 1		Problem 2				Problem 3		Memory
	Par 3	Par 4	Ordering		Answering		Par 3	Par 4	Test
			Par 3	Par 4	Par 3	Par 4			
p(correct evidence)	.033	.500	.333	.200	.300	.310	.034	.276	.429
p(wrong evidence)	.033	.567	.400	.333	.867	.759	.207	.724	.179
p(correct inference)	.000	.267	.033	.067	.433	.414	.034	.517	.357
p(wrong inference)	.000	.600	.167	.067	.900	.379	.000	.517	.071
p(consistent conjunct.)	.000	.033	.000	.000	.067	.034	.000	.103	.000
p(inconsist. conjunct.)	.000	.000	.033	.033	.033	.000	.000	.000	.000
Ambiguous	.700	2.033	.900	.733	1.400	1.310	.655	1.379	1.107
N	30	30	30	30	30	29	29	29	28

Combined Categories: Evidence/Inference.

	Problem 1		Problem 2				Problem 3		Memory
	Par 3	Par 4	Ordering		Answering		Par 3	Par 4	Test
			Par 3	Par 4	Par 3	Par 4			
p(evidence)	.067	1.067	.733	.533	1.167	1.069	.241	1.000	.607
p(inference)	.000	.867	.200	.133	1.333	.793	.034	1.034	.429
p(conjunction)	.000	.033	.033	.033	.100	.034	.000	.103	.000
Ambiguous	.700	2.033	.900	.733	1.400	1.310	.655	1.379	1.107
N	30	30	30	30	30	29	29	29	28

Table 23
Subjects' use of the conditional probability to which they were exposed, after paragraph 4 of the fixed-order problems.

Conditional probability given in each problem.	S type	Problem the subject was working on. Problem 1 Ave # of times S used concept Evid	Infer	Problem 3 Ave # of times S used concept Evid	Infer	
p(e/h) in Prob 1, p(h/e) in Prob 3	Undergr	1.00*	1.57	0.43	1.71*	n=7
	Grad st	1.27*	0.83	1.00	1.67*	n=6
	Total	1.12*	1.23	0.69	1.69*	n=13
p(h/e) in Prob 1, p(e/h) in Prob 3	Undergr	0.29	0.14*	1.00*	0.27	n=7
	Grad st	1.33	1.33*	2.17*	1.00	n=6
	Total	0.77	0.69*	1.54*	0.62	n=13

Note: Data are mean numbers of mentions of evidence and inference type conditional probabilities. * indicates the cells expected to have larger means. Each subject appeared in either the first or the second set of rows.

Table 24
Subjects' use of the conditional probability to which they were exposed, subject-selected order problems.

Order the conditional probability stimulus paragraphs were presented in.	Activity[a]	Paragraph the subject was referring to. Paragraph 3 Ave # of times S used concept: Evid	Infer	Paragraph 4 Ave # of times S used concept: Evid	Infer
p(e/h) third, p(h/e) fourth (n=12)	Ordering	0.67*	0.08	0.25	0.33*
	Answering	2.42*	0.83	0.50	1.17*
	Total	3.09*	0.91	0.75	1.50*
p(h/e) third, p(e/h) fourth (n=13)	Ordering	1.08	0.39*	0.92*	0.00
	Answering	0.46	1.54*	1.46*	0.69
	Total	1.54	1.92*	2.39*	0.69

Note: Data are mean numbers of mentions of the specified type of reliability concept.

*These cells were expected to have larger means. Each subject appeared in either the first or the second set of rows.

[a]Each subject both ordered and answered.

Table 25
Subjects' use of the conditional probability to which they were exposed, on fixed-order and subject-selected order problems and on memory test.

	Type of Conditional Probability Subject Spoke about			
Fixed-order problems.				
Type of Conditional Probability Given	Evidence	Inference	Mean	Proportion correct
Evidence p(e/h)	1.33*	0.93	1.13	.59
Inference p(h/e)	0.73	1.19*	0.96	.62
Mean	1.03	1.06	1.04	
Total proportion correct				.60
Subject-selected order problems.				
Evidence p(e/h)	2.73*	0.80	1.77	.77
Inference p(h/e)	1.16	1.72*	1.44	.60
Mean	1.95	1.26	1.61	
Total proportion correct				.69
Memory test.				
Evidence p(e/h)	0.61*	0.15	0.39	.80
Inference p(h/e)	0.60	0.67*	0.63	.53
Mean	0.61	0.43	0.52	
Total proportion correct				.62

Note: * indicates the cells expected to have larger means.

Appendix I. The three fixed-order problems used in this research, with p(e/h) conditional probability paragraphs.

Doctor Problem.

The next word problem is about a doctor trying to figure out what disease a patient has. The patient is clearly sick, but it is hard to know what disease he has. You will be asked to estimate how likely it is that the patient has one of two diseases.

Problem definition. The patient comes in to the emergency room at night with a very unusual symptom - his eyeballs are bright yellow. The doctor knows that there are only two diseases that can produce that symptom - hepatitis and toxic uremia. People never get them both at the same time.

*With what you know now, what is the probability that the patient has toxic uremia?*_____

Base rate. A discussion with a colleague reminds the doctor that toxic uremia is a less common disease than hepatitis. He checks a textbook and finds that 85% of people with the symptom of yellow eyes have hepatitis, and only 15% of them have toxic uremia.

*With what you now know, what is the probability that the patient has toxic uremia?*_____

Evidence. The doctor orders the lab to do a Spock test on the patient's blood. In two hours the results are back - the Spock test indicates that the patient has toxic uremia.

*With what you know now, what is the probability that the patient has toxic uremia?*_____

Reliability: p(e/h). The doctor consults his diagnostic manual and discovers that the Spock test is the best way to find out whether a patient with yellow eyes has hepatitis or toxic uremia. However, the Spock test is not perfect. It has an error rate of 20%, and is right 80% of the time. That is, when the patient has toxic uremia, the Spock test says so 80% of the time, but it falsely indicates that the patient has hepatitis 20% of the time. Similarly, when the patient has hepatitis, the Spock test will indicate that the disease is toxic uremia about 20% of the time.

*With what you know now, what is the probability that the patient has toxic uremia?*_____

Termination Problem.

Problem definition. Catherine Shelton, an auto claims adjuster for Sparta Insurance Company, is especially bothered by a claim that sits on her desk. This particular insurance claim concerns the insured, Kenneth Peterson, whose name is familiar to Miss Shelton. Since he moved to Colorado 10 months ago, upon his retirement from the Air Force (he had served nearly 20 years in Europe), he has had two auto accidents. The damages in the accident 4 months ago were relatively minor, but in the current accident someone was injured, and Mr. Peterson received a citation from the police. Miss Shelton wonders if this accident pattern for Peterson will continue. Perhaps, she reasons, Mr. Peterson is a bad risk for Sparta.

Miss Shelton has worked in claims adjustment for 5 years and knows, in terms of dollars and cents, what a bad insurance risk means to an insurance company. If the insured is a poor driver and, consequently, a bad risk for the company, then his insurance policy with the company should be terminated, or he should be moved from a "preferred" policy to a "standard" one with higher premiums. However, although the average driver has no accidents in any given year, or even in any five year period, still it is true that even very cautious and experienced drivers become involved in automobile accidents occasionally through no fault of their own. It really isn't right to penalize a consumer in those cases. These people are good customers, good risks, who probably will not have automobile accidents in the future, and if the agent terminates them or raises their rates (which will probably make them move to another company), it is a good customer lost.

*What is the likelihood that Kenneth Peterson is a good insurance risk, i.e., that he will not have an accident during the next five years?*_____

Base rate. To aid her thinking about whether Mr. Peterson is a good insurance risk for the Company, Miss Shelton consults actuarial data. She discovers that 65% of those people who have 2 or more accidents in a given year will have another accident during the 5 years which follow.

*With the information you have now, what is the likelihood that Kenneth Peterson is a good insurance risk, i.e., that he will not have an accident during the next five years?*_____

Evidence. Miss Shelton decides that she needs more information about Mr. Peterson's driving record. His application stated that he had had no accidents in the past 5 years. Since she does not know where he served in Europe, nor how to access official records concerning the driving records of armed services personnel, she does not have a way of verifying this information. She decides to telephone Mr. Peterson and ask him directly about his driving record. When she calls, he unhesitatingly responds that he had not had any driving accidents within the previous 15 years.

*With the information you have now, what is the likelihood that Kenneth Peterson is a good insurance risk, i.e., that he will not have an accident during the next five years?*_____

Reliability: p(e/h). Miss Shelton realizes that Mr. Peterson might have deceived her; however, she is also aware that when people are asked a surprising question, they usually answer truthfully. A study she read recently in the Journal of Insurance Psychology indicated that only 25% of people who had poor driving records attempted to hide this when called by someone who identified himself as a representative of their insurance company.

*With all the information you have now, what is the likelihood that Kenneth Peterson is a good insurance risk, i.e., that he will not have an accident during the next five years?*_____

Twins Problem.

Problem definition. The next word problem is about two boys. One of them broke a lamp. You will not know for sure which one did it. You will be asked to estimate the probability that one of them was the one who did it.

Stephen and Paul are 5 year old twins. One afternoon their mother hired a new babysitter so she could go out to do errands. Before she left, she took the sitter aside and gave her some advice about handling the boys.

*With what you know now, what do you think is the probability that Stephen is the one who broke the lamp?*_____

Base rate. On her way out, the twins' mother had told the babysitter: "Paul is usually the troublemaker: I'd say about 80% of the time if one of them breaks a rule or does something careless, it is Paul."

*With what you know now, what do you think is the probability that Stephen is the one who broke the lamp?*_____

Evidence. The sitter was preparing a snack in the kitchen. When she glanced into the living room to check on the boys, she saw one of them, she thought it was Stephen, standing half on the couch and half on the lamp table, reaching for something on a shelf. Before she could turn off the water and come out to make him stop, she heard a crash. Running to the living room, she found Stephen and Paul and a broken lamp. She asked Stephen, "Did you knock over the lamp?" "No", he answered, "Paul did." But Paul shouted, "No, Stephen did it."

With what you know now, what is the probability that it was Stephen who broke the lamp?
_____

Reliability: p(e/h). Stephen's and Paul's mother enjoys dressing them alike. Before she left, she had said to the babysitter "New people have trouble telling the boys apart. I'd say they only identify them correctly 60% of the time. So two times in five, when it is Stephen, you think it is Paul, or if it is really Paul, you think it is Stephen."

*With what you know now, what do you think is the probability that Stephen is the one who broke the lamp?*_____

Appendix II. The alternative conditional probability paragraphs, with p(h/e) statements, for all three problems.

Doctor problem.

The doctor consults his diagnostic manual and discovers that the Spock test is the best way to find out whether a patient with yellow eyes has hepatitis or toxic uremia. However, the Spock test is not perfect. It has an error rate of 20%, and is right 80% of the time. That is, when the Spock test says the patient has toxic uremia, the patient actually has it 80% of the time, but the patient has hepatitis 20% of the time. Similarly, when the Spock test says that the patient has hepatitis, the patient will actually have toxic uremia about 20% of the time.

Insurance problem.

Miss Shelton realizes that Mr. Peterson might have deceived her; however, she is also aware that when people are asked a surprising question, they usually answer truthfully. A study she read recently in the Journal of Insurance Psychology indicated that only 25% of people who said that they had good driving records, when called by someone who identified himself as a representative of their insurance company, were actually hiding a poor driving record.

Twins problem.

Stephen's and Paul's mother enjoys dressing them alike. Before she left, she had said to the babysitter "New people have trouble telling the boys apart. I'd say they only identify them correctly 60% of the time. So two times in five, when you think it is Stephen, it is really Paul, or if you think it is Paul, it is really Stephen."

Appendix III. The subject-selected order version of the Twins Problem.

Problem definition. The next word problem is about two boys. One of them broke a lamp. You will not know for sure which one did it. You will be asked to estimate the probability that one of them was the one who did it.

Stephen and Paul are 5 year old twins. One afternoon their mother hired a new babysitter so she could go out to do errands. Before she left, she took the sitter aside and gave her some advice about handling the boys.

With what you know now, what do you think is the probability that Stephen is the one who broke the lamp?_____

The following paragraphs show kinds of information that is available for you to use in answering the question on the previous page. You will be given the information that is blanked out, in the order that you request it. You will answer the question again after each paragraph is filled in.

Please indicate the order in which you would like to receive the information. That is, put a "1" below the paragraph whose blanks you would like to have filled in first, then a "2" below the one you would like to receive next, then a "3" and a "4". No ties are allowed, and please do not leave any paragraphs blank.

1. **Base rate.** On her way out, the twins' mother had told the babysitter: "_____ is usually the troublemaker: I'd say about 80% of the time if one of them breaks a rule or does something careless, it is _____."

   Order _____ _____

2. **Evidence.** The sitter was preparing a snack in the kitchen. When she glanced into the living room to check on the boys, she saw one of them, she thought it was _____, standing half on the couch and half on the lamp table, reaching for something on a shelf. Before she could turn off the water and come out to make him stop, she heard a crash. Running to the living room, she found both boys and a broken lamp. Each of them claimed the other broke it.

   Order _____ _____

3. **Reliability: p(e/h).** Stephen's and Paul's mother enjoys dressing them alike. Before she left, she had said to the babysitter "New people have trouble telling the boys apart. I'd say they only identify them correctly ____% of the time. So ____ times in five, when it is Stephen, you think it is Paul, or if it is really Paul, you think it is Stephen."

   Order _____ _____

4. **Reliability: p(h/e).** Stephen's and Paul's mother enjoys dressing them alike. Before she left, she had said to the babysitter "New people have trouble telling the boys apart. I'd say they only identify them correctly ____% of the time. So ____ times out of five, when you think it is Stephen, it is really Paul, or if you think it is Paul, it is really Stephen."

   Order _____ _____

Appendix IV. Excerpts from research assistants' session script.

Procedure. The subject will be dealt with as follows:

1. Meet, greet, and welcome them.
2. Explain that it is a study of how they solve problems that they have probably not previously been taught how to do. Tell them that they will be asked to explain their thinking.
3. Give them the first sheet, "Information about the Study".
4. Fill out the "Researcher Notes" sheet with name, date, and the code number of the question booklet. Use this sheet to note anything unusual about the procedure, or anything interesting that you observe.
5. Read the Practice Thinking Aloud sheet, and do a practice think-aloud problem. Research say, "If I nod and smile as you think aloud, it is only to encourage your thinking aloud; I will *not* be telling you whether your answer is right, or whether you are on the right track."
6. Give them the booklet and have them fill out the top sheet.
7. First problem (order is determined by the order the problems appear in the booklet). The procedure will be:
   a. First, give them their particular think aloud instructions: show them the sheet of paper with "in your own words, tell me what the problem is, what information you have available, and how you solve the problem". This is to be their guide for each paragraph.
   b. Turn to the problem, cover it with a sheet of paper (it doesn't matter whether you turn the pages and cover them for the subject, or just tell them what to do)
   c. Turn on the tape recorder, and identify Subject, Date, Study (Think aloud study number 2), and researcher. Note the time: for you have a half hour of tape (at high speed).
   d. Expose just one paragraph at a time (researcher will say aloud "I am now exposing the (second, third, fourth) paragraph" when the sheet is moved, or else have the subject say it).
   e. Have the subject read the paragraph silently,
   f. then say "okay" and read the question aloud
   g. and answer it thinking aloud, guided by the "Think Aloud Instructions" sheet. Tell the subject that he/she is allowed to look back at the paragraph, and at previous paragraphs, in answering the probability questions and in coming up with an explanation, but is not to simply read the material (or selections) aloud again; that would not be an acceptable explanation.
      i. If the subject falls silent before writing his or her answer, prompt with "keep talking".
      ii. If the researcher thinks of a focussed question to ask, it should be written down. At the end, after the memory test, you can take the subject back to the problem, get them to review the context, and ask the question.
   h. Subject should write his/her answer in the blank.
   i. After the first paragraph *and* the third paragraph, have the subject read the Think Aloud Instructions again. [Do this on the third problem as well.]
   j. The same procedure is to be followed for each of the four paragraphs.
8. Second problem. In this problem the subject will indicate which information he or she wants.
   a. Subject reads first paragraph, estimates a probability, following same think-aloud instructions as for the first problem.
   b. Give subject instructions for thinking aloud about second page of this problem:
      i. For the rest of this problem, I want you to read everything aloud and think aloud as you follow the instructions. Tell me every idea that comes to mind.

This applies during both the *ordering* of the paragraphs, and then the *answering of the question* after the researcher fills in the blanks in each paragraph in turn.

c. Turn to second page. Subjects
   i. read top paragraph of second page, aloud
   ii. go back and read *question* on previous page. [explain to them that the point of ordering the information on this page is to find the right answer to that question, as soon as possible]
   iii. [tell them that they may refer back to the paragraph on that previous page any time they want to]
   iv. read the four paragraphs with blanks, aloud; thinking aloud as they go and saying how useful that information would be
   v. go back and read the question again
   vi. Order the four paragraphs. Ask subject, "Why would this information be valuable for answering the question?", each time they specify a rank order for one of the paragraphs.

d. and then specify the order in which they want to get the information,

e. and think aloud while making these decisions. Researcher prompts them if they fall silent. Since this involves reading, have them read aloud.

f. Subject is given the missing information (information is on a sheet of paper in the researcher's materials), in the requested order, and answers the p(H) question after each of the four pieces, thinking aloud while doing it.

9. Third problem. As with the first.

10. Surprise Memory Test - Researcher uses the sheet to ask the subjects to explain their thinking on the first problem, again, from memory, without referring back to it. Researcher refreshes them on the topic and the question. They have to remember the information and the logic of their answer.

11. Debriefing. Tell the subject the point of the study. There is nothing deceptive here. We are studying how people solve this kind of word problem. Give them the second information handout, "Feedback for Psych 100 students".

12. Give them their filled out yellow card (fill it out earlier if you get a chance), thank them for helping science.

13. Record name and date on the cassette.

14. Staple the Researcher Notes to the front of the booklet.

Practice Thinking Aloud

[Researcher reads this aloud to the subject:]

The study in which you are participating is being conducted to determine how people think about word problems. In this session you will be given three written word problems asking you to estimate the probability that something is true. I would like you to read each problem and, on the basis of your own knowledge and the information you read, to answer every question.

I want you to think aloud while you answer the word problems. That is, say EVERYTHING you are thinking, including, for example, visual images, ideas you have for possible strategies for answering the questions, and whatever answers you are thinking about, even if you are not sure they are right.

I want you to speak spontaneously and constantly. If you fall silent I will prod you to tell me what you are thinking. I will be taping what you say.

[Practice example, such as "How many windows are there in the house that you grew up in", or "Multiply in your head: 57 times 23".]

I will give you special instructions on each of the problems.

Memory Test.

When the subject has finished with all three problems, it is time for a surprise question:

"I would now like you to tell me about the first problem that we did today. I will ask you four questions, and tell me everything you can remember about each one. The problem was about (Twins, Doctors, Insurance), and I want you to tell me in your own words what the problem was. [First question.]

[After they answer, tell them: "The exact question was, in fact,

1. [twins]: what do you think is the probability that Stephen is the one who broke the lamp?
2. [insurance]: What is the likelihood that Kenneth Peterson is a good insurance risk, i.e., that he will not have an accident during the next five years?
3. [doctor]: What is the probability that the patient has toxic uremia?

The Memory Test Questions:

Please answer these questions: [Remember to let subject answer each fully before asking the next.]

1. [Second question.] What information given in the four paragraphs of the first problem was important for you in answering the question?
2. [Third question.] What was the final answer that you gave to the question?
3. [Fourth question.] How did you get your answer?

After they are fully done, you can give them feedback, tell them whether they remembered their answer, or whatever else they are curious about.

Think Aloud Instructions.

The word problems are divided into separate paragraphs, each followed by a question. I would like you to read each **paragraph** silently, and then say "Okay" and read the **question** aloud, and then:

IN YOUR OWN WORDS, TELL ME

THE PROBLEM THAT YOU HAVE TO SOLVE HERE,

THE INFORMATION YOU HAVE AVAILABLE TO USE,

AND HOW YOU SOLVE THE PROBLEM.

You are free to look back at the paragraph while you are thinking, but please *do not read the paragraph aloud again.*

On later paragraphs on the page, you are allowed to look up at the earlier paragraphs to remind yourself about needed information.

Numbers for the Second Problem

[Researcher - keep this information hidden!]

1. If the second problem is Doctors, the information for the paragraphs is:
   a. 85% 15%
   b. toxic uremia
   c. 80% 20% 20%
   d. 80% 20% 20%
2. If the second problem is Insurance, the information for the paragraphs is:
   a. 65%
   b. zero
   c. 25%
   d. 25%
3. If the second problem is Twins, the information for the paragraphs is:
   a. Paul Paul
   b. Stephen
   c. 60% 2
   d. 60% 2

Appendix V. Definition of conditional probability categories for each of the three problems.

Doctor Problem.

1. p(h/e). Doctor: probability of having toxic uremia given Spock test says it was toxic uremia.
2. p(h/~e). Doctor: probability that the disease is toxic uremia given that the test says it is hepatitis.
3. p(~h/e). Doctor: probability that the disease is hepatitis given that the Spock test says it is toxic uremia.
4. p(~h/~e). Doctor: probability of having hepatitis given that the Spock test says he has hepatitis.
5. p(e/h). Doctor: probability that the test will say the patient has toxic uremia given that the patient has toxic uremia.
6. p(e/~h). Doctor: probability of the test saying the patient has toxic uremia given that the patient has hepatitis.
7. p(~e/h). Doctor: probability of the test saying the patient has hepatitis given that the patient has toxic uremia.
8. p(~e/~h). Doctor: probability of the test saying someone has hepatitis given that they do have hepatitis.
9. p(correct evidence), i.e., p(e/h) or p(~e/~h). Doctor: probability of the test correctly identifying the disease.
10. p(wrong evidence), i.e., p(e/~h) or p(~e/h). Doctor: probability of the test being wrong, saying the person has the wrong disease.
11. p(correct inference), i.e., p(h/e) or p(~h/~e). Doctor: probability of the patient having the disease that the Spock test says he has.
12. p(wrong inference), i.e., p(h/~e) or p(~h/e). Doctor: probability of the patient having the other disease from what the test said they had.
13. p(h and e) [or p(e and h)]. Probability of having toxic uremia and having a "toxic uremia" result on the Spock test.
14. p(h and ~e) [or p(~e and h)]. Probability of having toxic uremia but having a "hepatitis" result on the Spock test.
15. p(~h and e) [or p(e and ~h)]. Probability of having hepatitis but having a "toxic uremia" result on the Spock test.
16. p(~h and ~e) or [p(~e and ~h)]. Probability of having hepatitis and having a "hepatitis" result on the Spock test.
17. They discuss the concept of reliability but it is difficult to understand which of the above concepts they were using.
18. They did not mention any reliability or conditional probability concept.

Insurance Problem.

1. p(h/e). Insurance/termination: probability the client had a good past driving record given he says he had a good past driving record.
2. p(h/~e). Insurance/termination: probability that the client had a good past driving record given that he admits to a bad past driving record.
3. p(~h/e). Insurance/termination: probability that the client had a bad past driving record given that he said on the phone that he had a good past driving record.
4. p(~h/~e). Insurance/termination: probability that the client had a bad past driving record given that he said on the phone that he had a bad driving record.
5. p(e/h). Insurance/termination: probability a client will say he has a good past driving record given that he or she has a good past driving record.
6. p(e/~h). Insurance/termination: probability of a client saying he or she has a good past driving record given that he or she has a bad past driving record.
7. p(~e/h). Insurance/termination: probability of someone admitting to having a bad past driving record given that they had a good past driving record.
8. p(~e/~h). Insurance/termination: probability of someone admitting to a bad past driving record given that they have a bad past driving record.
9. p(correct evidence), i.e., p(e/h) or p(~e/~h). Insurance/termination: probability of someone truthfully reporting on their driving record.
10. p(wrong evidence), i.e., p(e/~h) or p(~e/h). Insurance/termination: probability of someone mistakenly reporting, or lying, about their driving record (Note that the motivations for these two would be different).
11. p(correct inference), i.e., p(h/e) or p(~h/~e). Insurance/termination: probability that the person had the kind of driving record that they said they had.
12. p(wrong inference), i.e., p(h/~e) or p(~h/e). Insurance/termination: probability of someone having a different driving record from what they report they had.
13. p(h and e) [or p(e and h)]. Probability that a person would both have a good driving record and say that they had a good driving record.
14. p(h and ~e) [or p(~e and h)]. Probability that a person would have a good driving record but say that they had a bad driving record.
15. p(~h and e) [or p(e and ~h)]. Probability that a person would have a bad driving record but say that they had a good driving record.
16. p(~h and ~e) or [p(~e and ~h)]. Probability that a person would both have a bad driving record and say that they had a bad driving record.
17. They discuss the concept of reliability but it is difficult to understand which of the above concepts they were using.
18. They did not mention any reliability or conditional probability concept.

Twins Problem.

1. p(h/e). Twins: probability that the twin who broke the lamp was Stephen given that the sitter thought she saw Stephen.
2. p(h/~e). Twins: probability that Stephen was the twin who broke the lamp given that the sitter thought she saw Paul.
3. p(~h/e). Twins: probability that Paul was the twin who broke the lamp given that the babysitter thought she saw Stephen.
4. p(~h/~e). Twins: probability that Paul was the twin who broke the lamp given that the babysitter thought she saw Paul.
5. p(e/h). Twins: probability of someone saying they saw Stephen given that it was Stephen.
6. p(e/~h). Twins: probability of someone thinking they see Stephen given that it was really Paul.
7. p(~e/h). Twins: probability of someone thinking that they saw Paul given that it was really Stephen.
8. p(~e/~h). Twins: probability of someone thinking they saw Paul given that it really was Paul.
9. p(correct evidence), i.e., p(e/h) or p(~e/~h). Twins: probability of someone accurately identifying a twin.
10. p(wrong evidence), i.e., p(e/~h) or p(~e/h). Twins: probability of someone misidentifying one of the twins.
11. p(correct inference), i.e., p(h/e) or p(~h/~e). Twins: probability that the twin really is the one that the babysitter thought she saw.
12. p(wrong inference), i.e., p(h/~e) or p(~h/e). Twins: probability of it being the other twin from the one the babysitter thought she saw.
13. p(h and e) [or p(e and h)]. Probability that the babysitter would have thought Stephen broke the lamp and Stephen would have actually been the one that broke the lamp.
14. p(h and ~e) [or p(~e and h)]. Probability that the babysitter would have thought Paul broke the lamp but Stephen would have actually been the one that broke the lamp.
15. p(~h and e) [or p(e and ~h)]. Probability that the babysitter would have thought Stephen broke the lamp but Paul would have actually been the one that broke the lamp.
16. p(~h and ~e) or [p(~e and ~h)]. Probability that the babysitter would have thought Paul broke the lamp and Paul would have actually been the one that broke the lamp.
17. They discuss the concept of reliability but it is difficult to understand which of the above concepts they were using.
18. They did not mention any reliability or conditional probability concept.

Notes

[1]Medin and Edelson (1988) have identified a distinct mechanism that also is capable of producing neglect of a relative frequency that the subject has experienced.

[2]On the subject-selected order problem and the memory test, a higher proportion of the subjects' conditional probability concepts were appropriate when *evidence* conditional probabilities were presented (.77 and .80, respectively) than when *inference* conditional probabilities were presented (.60 and .53). This is consistent with a finding of Tversky and Kahneman's (1980). They "asked subjects to compare the two conditional probabilities p(Y/X) and p(X/Y) for a pair of events X and Y such that (1) X is naturally viewed as a cause of Y; and (2) p(X) = p(Y), that is, the marginal probabilities of the two events are equal. The latter condition implies that p(X/Y) = p(Y/X). Our prediction was that most subjects would view the causal relation as stronger than the diagnostic relation, and would erroneously assert that p(Y/X) > p(X/Y)" (p 119). If we view their "causal relation" as our *evidence* type conditional probability, p(e/h), and their "diagnostic relation" as our *inference* conditional probability, p(h/e), then the results of the two studies are coherent: we find that people understand and use the causal (evidence) information when they have it, more than they understand and use the diagnostic (inference) information; Tversky and Kahneman find that people think the causal connection is larger than the diagnostic one.