Diagnostic Inference: People's Use of Information in Incomplete Bayesian Word Problems.

Robert M. Hamm. PhD

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

August 1987

Institute of Cognitive Science
Publication Number 87-11

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
DIAGNOSTIC INFERENCE: PEOPLE'S USE OF INFORMATION IN INCOMPLETE BAYESIAN WORD PROBLEMS.

5. TYPE OF REPORT & PERIOD COVERED

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Robert M. Hamm

8. CONTRACT OR GRANT NUMBER(s)
MDA-903-86-K-0265

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Institute of Cognitive Science
University of Colorado, Box 345
Boulder CO 80309-0345

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
2Q161102B74F

11. CONTROLLING OFFICE NAME AND ADDRESS
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, Virginia 22333-5600

12. REPORT DATE
August 1987

13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS (of this report)
Unclassified

15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES
Contracting Officer's Representative was

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Inference, Probability, Probabilistic Inference, Heuristic strategies, Information Integration, base rate fallacy

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
Probabilistic inference word problems require people to integrate three types of information concerning a hypothesized cause of an event -- base rate information p(H) concerning the relative frequency of the cause in question, evidence E that the cause was responsible, and the reliability of that evidence p(E/H) -- and to evaluate the probability p(H/E) that the cause in question was responsible for the event. Three classes of hypothesis are proposed to explain how people answer these word problems and why the answers often neglect base rate

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Box 20, continued.

information -- normative probabilistic reasoning, heuristic strategies, and non-normative information integration. In a questionnaire study, 265 students estimated the probability of the cause before and after each type of information was presented. Information was presented in 6 orders, so some subjects responded to each possible subset of the information. Findings include: many subjects respond with numbers that are available in the problem presentation; the more recent information has a greater impact; there is no universally applied weighted averaging scheme that accounts for the average response in all conditions; the typical subject's responses are well described in terms of the use of strategies contin gent on the kind of information that is available. A production system simulation of the typical responses and the normative responses shows that the neglect of the base rate information is due in part to a misunderstanding of the reliability information, specifically, a confusion between p(H/E) and p(E/H). Recommendations for improving probabilistic inference are evaluated in the light of these results.

Table of Contents

Abstract. 0
1. Introduction 1
2. Previous research on probabilistic inference word problems. 1
2.1. People's strategies for solving probabilistic inference word problems. 3
2.2. The psychological process: Hypotheses about the subjects' strategies. 4
3. Methods. 7
4. Results. 7
4.1. Accuracy. 8
4.2. Evaluation of the hypotheses. 8
4.3. Hypotheses involving variants of normative probabilistic reasoning. 8
4.4. Hypotheses involving the non-normative integration of the available information. 9
4.4.1. The use of available numbers. 9
4.4.2. Application of arithmetic operations to available numbers. 10
4.4.3. Ambiguity of strategy identification. 11
4.5. Weighted average of information. 13
4.5.1. The integration condition. 14
4.5.2. The betweenness condition. 14
4.5.3. The nearness condition. 15
4.5.4. The universality of weights condition. 15
4.5.5. Tests of position dependent information weighting patterns. 15
4.5.6. Tests of content dependent information weighting patterns. 16
4.5.7. Other composition principles. 18
4.6. Simplifying and heuristic strategies. 19
4.6.1. Contingency in the neglect of information. 19
4.6.2. Increased accuracy due to contingent strategy use. 21
4.6.3. Variation in accuracy of the same strategy in different situations. 21
4.6.4. Conclusions concerning heuristic strategies. 26
4.7. Summary of the tests of the hypotheses. 26
4.8. Incidental Results. 27
4.8.1. Differences between problems. 27
4.8.2. Subject stability over problems. 28
4.8.3. Subject factors that influence accuracy and strategy choice. 30
4.8.4. Accuracy of forced choice. 31
5. Discussion. 32
5.1. Significant findings. 32
5.2. Implications of the neglect of base rates in word problems. 34
6. Bibliography. 36
7. Tables. 40

Diagnostic Inference: People's Use of Information in Incomplete Bayesian Word Problems.

Abstract.

Probabilistic inference word problems require people to integrate three types of information concerning a hypothesized cause of an event -- base rate information p(H) concerning the relative frequency of the cause in question, evidence E that the cause was responsible, and the reliability of that evidence p(E/H) -- and to evaluate the probability p(H/E) that the cause in question was responsible for the event. Three classes of hypothesis are proposed to explain how people answer these word problems and why the answers often neglect the base rate information -- normative probabilistic reasoning, heuristic strategies, and non-normative information integration. In a questionnaire study, 265 students estimated the probability of the cause before and after each type of information was presented. Information was presented in 6 orders, so some subjects responded to each possible subset of the information. Findings include: many subjects respond with numbers that are available in the problem presentation; the more recent information has a greater impact; there is no universally applied weighted averaging scheme that accounts for the average response in all conditions; the typical subject's responses are well described in terms of the use of strategies contingent on the kind of information that is available. A production system simulation of the typical responses and the normative responses shows that the neglect of the base rate information is due in part to a misunderstanding of the reliability information, specifically, a confusion between p(H/E) and p(E/H). Recommendations for improving probabilistic inference are evaluated in the light of these results.

Diagnostic Inference. August 11, 1987
Robert M. Hamm, University of Colorado.

1. Introduction

It is possible to measure the extent to which one believes a proposition. Probability is a recommended measure, both because most people are familiar with it, and because it has some very convenient mathematical features (Krantz, Luce, Suppes, and Tversky, 1971). For example, if one believes proposition H with probability p(H), one should believe the contradiction of H with probability 1 - p(H) (complementarity); and if one already has a degree of belief p(H) in proposition H, and one is given new evidence E pertinent to the truth of H, one can use a specific formula (Bayes' Theorem) to adjust one's degree of belief in H to a new p(H/E). If people would use probabilities correctly to measure their degrees of belief, they could communicate their uncertainties accurately (see Wallsten, Budescu, Rapoport, Zwick, and Forsyth, 1986) and they could consider the expected utilities of their decision options (Edwards, 1961).

Research has shown, however, that people's subjective probabilities are not precise (Slovic, Fischhoff, and Lichtenstein, 1977), even with some training (Lichtenstein and Fischhoff, 1980). For example, with word problems that present information pertinent to the establishment and subsequent revision of the degree of belief in a hypothesis, people do not use the information properly (Fischhoff and Bar-Hillel, 1984; Kahneman and Tversky, 1972; Tversky and Kahneman, 1982.). Typically, they neglect the information pertinent to the establishment of the belief.

The present study uses a new method to study changes in the subjects' degree of belief in the hypothesis while reading and answering a probabilistic inference word problem. Instead of presenting the information in one standard order (first the base rate information which establishes the probability of the hypothesis, then the evidence pertaining to the hypothesis, and finally the reliability of that evidence) and then asking for the subject's final degree of belief in the hypothesis, this method presents the information in each possible order (to different subjects) and asks for the degree of belief before and after each piece of information. This design allows

1. the investigation of people's probabilistic inferences in situations which have norms other than Bayes' Theorem,
2. the testing of a number of hypotheses concerning the process by which people produce and change their probabilistic degree of belief in a hypothesis, during word problems.

2. Previous research on probabilistic inference word problems.

A probabilistic inference word problem is a word problem that presents a situation and gives information relevant to the reader's degree of belief in various propositions. The information is presented either as observations (evidence), as numerical probabilities, as relative frequencies, or as verbal expressions of probability. The subject is asked the probability of a proposition about the problem situation. Subjects' behavior is described and compared with the normative use of the information to answer the question. The norm we are primarily concerned with here, Bayes' Theorem, can be applied when one has a degree of belief in a hypothesis, p(H) (the "prior probability"), one knows the probability of observing a particular piece of evidence if the hypothesis is true, p(E/H), and also if the hypothesis is false, p(E/~H) (the "conditional probabilities"), and one observes evidence E. In word problems with this information, p(H) should be revised to a "posterior probability", p(H/E), according to the following rule:

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

The first wave of research on how people do word problems where Bayes' Theorem is the norm was the "book bag and poker chip" paradigm (see Edwards, 1968; Slovic and Lichtenstein, 1971). The typical word problem is given in Table 1. These problems were often encountered in a laboratory, where the evidence was an event that happened "in the present" rather than being read about; thus it was a particularly vivid "word problem". Subjects usually answered a number of questions about each word problem, following the presentation of new information. Usually a number of chips would be drawn, and the subject would report the probability that it was the predominantly red bag after each drawing. The subjects were given the following sequence of

information in this problem: (a) p(E/H) = .7 and p(E/~H) = .3; (b) p(H) = .5; (c) a series of pieces of evidence, each either E or ~E. They answered p(H/all E so far presented) after each E. [Recent examples, such as Lopes (1982) and Robinson and Hastie (1985), have presented a series of pieces of evidence E_i about which the reliability $p(E_i/H)$ was not uniform for all i, and was not given explicitly.]

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

A common finding in these studies is "conservatism", that subjects did not adjust their degree of belief in H as far as Bayes' Theorem would prescribe. This was ascribed both to their failure to appreciate how rare it would be to draw a particular combination of chips from one of the bags ("misperception"), and to their failure to combine the information properly ("misaggregation") (see Edwards, 1968).

The second wave of research (see Fischhoff and Bar-Hillel, 1984; Tversky and Kahneman, 1982) used word problems like the Cab problem (see Table 2). This research commonly used questionnaire studies in which the subject answered only one problem of this type. Here the problem indicates that the prior probability p(H) is much less than .5. The prior probability is not stated directly, but rather a relative frequency or base rate is given, which the subject must recognize to be an estimate of the prior probability. There is only one piece of evidence, which favors the unlikely hypothesis H. The sequence of information a subject encounters in the Cab problem is: (a) p(H) is .15; (b) E (favoring H) was observed; (c) p(E/H) = .8 and p(E/~H) = .2. The subject answers p(H/E) after reading all this information.

****
Insert Table 2 about here.
****

The typical finding with the Cab problem and its variants is that the p(H) information is ignored or insufficiently attended, when compared with the Bayes' Theorem prescription. Hence, the evidence is given *too much* weight, the opposite of the conservatism found with the Bookbag problem (Bar-Hillel, 1980). It should be noted, however, that in problems of the Cab type only one piece of evidence is given, and a similar nonconservatism had been witnessed with Bookbag problems following the *first* piece evidence in the series (Peterson and Miller, 1965; see Slovic and Lichtenstein, 1971, p 697). Additionally, the p(H) of .5 used with the Bookbag problem is almost "null" information (i.e., it is identical to the assumption subjects make if no p(H) is given). Some first wave researchers gave subjects p(H) values other than .5, and found an underutilization of this information (Slovic and Lichtenstein, 1971, p 703). Despite these links between the research findings using the two word problems, the behavior observed on the Cab problem has been seen as an instance of a different phenomenon, the "neglect of the base rate" (Tversky and Kahneman, 1982), which is also observed with word problems that give multiple, non-independent pieces of evidence (Kahneman and Tversky, 1973; Borgida and Brekke, 1981).

Research with variants of the Cab problem has explored the generality of the neglect of the base rate. For example, Cascells, Schoenberger, and Grayboys (1978) and Eddy (1982) observed it with doctors answering realistic medical word problems. A second approach to generalization has been to seek to influence the subjects' utilization of the prior probability information by either (a) varying the wording of the information pertinent to p(H), to see if subjects will then be more likely to use it (e.g., Bar-Hillel, 1980; Tversky and Kahneman, 1980), or (b) instructing subjects to pay attention to the base rate (e.g., Fischhoff, Slovic, and Lichtenstein, 1979; Fischhoff and Bar-Hillel, 1984; Lichtenstein and MacGregor, 1984). A general principle explaining variations in subjects' use of the base rate information has been offered by Bar-Hillel (1980, p 230): "People integrate two items of information only if both seem to them equally relevant.... One item of information is more relevant ... than another if it somehow pertains to it more specifically."

This principle explains a number of the experimental findings, including extreme cases such as

the presentation of only the base rate (the only information available, it is used; Tversky and Kahneman, 1982), and the presentation of no base rate information (subjects do not notice it is missing; Hammerton, 1973). The principle also covers the findings of studies that present subjects with a series of word problems over which the base rate information is varied (Birnbaum and Mellers, 1978; Fischhoff, Slovic, and Lichtenstein, 1979; see also Christensen-Szalanski and Bushyhead, 1981). Subjects respond to variations in the base rate, even if it is in fact irrelevant to the question (Fischhoff and Bar-Hillel, 1984; Fischhoff, Slovic, and Lichtenstein, 1979). The relevance principle accounts for this if we assume that the variation in the base rate between problems makes the subject think it is pertinent to the problems.

The relevance principle has been well received (von Winterfeldt and Edwards, 1986). It can give heuristic guidance in our attempts to change word problems, as well as real life probabilistic inference situations, so that people will use base rate information. However, it verges on tautology (defining "relevance" in terms of what makes people use base rate information on word problems). A more specific definition of psychological "relevance", and its determinants and consequences, is needed. Background for this sharper definition can be provided by a more detailed description of the processes by which people produce answers to word problems, which is the aim of the present study.

2.1. People's strategies for solving probabilistic inference word problems.

In previous research there has not been enough variation in the word problems to allow researchers to tease out the processes people use to answer the questions. First, only fairly difficult probabilistic inference problems (those for which Bayes' Theorem provides the right answer) have been studied, so we know little about whether people are able to make simpler inferences, or how they might do so. Second, while the Cab problem has three key pieces of information (see below), subjects have given their answer only after getting all three pieces, and so it is not known how each piece is used individually, how they are used in combination, nor whether there are any effects of the order of presentation.

The present study, which uses the standard Cab problem and two variants, identifies the three key pieces of information in each problem and presents them separately. Thus, the base rate or prior probability information, the evidence, and the reliability of the evidence are presented one at a time, in each of the six possible orders (for different subjects), and the subjects are asked for p(H) [or, if E has already been given, for p(H/E)] both before any information is given and after each piece. This allows us to

1. study the subjects' use of evidence and reliability information, in addition to base rate information;
2. compare subjects' performance with the norm, in a number of situations that have not previously been studied;
3. determine whether the order of reading the information influences the answer.

The method has been partially anticipated. Tversky and Kahneman (1982) review a study in which only the base rate information was given. They found that it was used (see also Locksley and Stangor, 1984). The present study extends this by giving the reliability alone, and the evidence alone. Fischhoff and Bar-Hillel (1984) required subjects to think about the problem using "Isolation Analysis", in which the subjects were presented with the full Cab problem and then asked:

1. "If you only knew the proportion of Green cabs in the city, what would you think is the probability that the cab was Green?"
2. "If you only knew the witness' reliability, what would you think is the probability that the cab was Green, as the witness claimed?"

These questions are analogous to (1) asking p(H) after only the base rate information is presented, and (2) asking p(H/E) after only the evidence and reliability are given. The present technique is better for our purposes because it is direct, while Isolation Analysis (which was intended as a

focusing technique) gives people information and then asks them to pretend that they do not have it while they answer a question, a procedure subject to hindsight bias (Fischhoff, 1975).

The study seeks a general account of how people answer probabilistic inference word problems. Their task is viewed as the production of a judgment from a variable number of inputs. Will it be possible to find an explanation of subjects' answers at every step of the problem, that also explains how people do the final step, the Bayesian inference problem that has been studied before?

The study also seeks a general description of people's accuracy in using probabilities to measure their degree of belief (in comparison with the norms of probability theory) on all the subproblems encountered in the present method. Each subproblem has either a normative answer, or a range of normatively acceptable answers. Will it be possible to find a general account of subjects' accuracy on these problems, that also accounts for how well they do on the Bayes' Theorem problems?

2.2. The psychological process: Hypotheses about the subjects' strategies.

The strategy of this study is to use the data to eliminate or support hypotheses about the processes by which people produce their answers to probabilistic inference word problems. The following hypotheses are listed in approximate order of increasing sophistication and understanding involved in the process.

1. **Non-normative integration of available numerical information.** People answer word problems by mentally combining the numerical information in some manner, without necessarily using the probabilities *qua* probabilities. The terms of Bar-Hillel's (1980) relevance hypothesis (quoted above) suggest this form of model (see also Einhorn, 1985).
   a. **Using one available number.** The subject may respond with one of the numbers presented explicitly or implicitly in the word problem. The selection of the number may not be guided by the meanings of the probability numbers; it may be random (Hamm, 1987).
   b. **Simple mathematical operations.** The subject may apply simple mathematical operations to the numbers in the word problem, such as complementation, addition, subtraction, multiplication, division, or combinations of these. For example, Kahneman and Tversky (1972, p 448) found that subjects in a variant of the book bag and poker chip word problem frequently used the sample ratio (the proportion of red chips in the sample) as an estimate of the probability that the "predominantly red bag" was the source.
   c. **Complex mathematical operations.** The subject may use more complicated operations. Krantz and Tversky (1971) list a number of possible combinations of three input variables A, P, and U: (A + P)*U, A*U + P, and A*P*U (and permutations). Variations of Bayes' formula provide another set of possibilities: (A*P)/(A*P + (1-A)*(1-P)).
   d. **Use of conventional probabilities.** The subject may use common landmarks on the probability line -- thirds, fourths, fifths, tenths -- either to express a global evaluative judgment, a simultaneous weighted average, or a sequential anchoring and adjustment process (see for example Kahneman and Tversky, 1972, p 447).
   e. **Weighted average of information.** The subject may integrate the available numerical information using some form of weighted average. Averaging is a common form of information integration (Anderson, 1981; Dawes, 1979; Hammond, Hamm, Grassia, and Pearson, in press) and has been suggested as an explanation of people's behavior in the Bookbag problems (Marks and Clarkson, 1972; Shanteau, 1972). This integration process is assumed to be applied universally; that is, the same relative weights are applied to any two pieces of available information, no matter

whether other information is present. Various weighted averaging theories may be distinguished by the relative weights put on information.

i. **Sequence-dependent information weighting pattern.** The weight the subject puts on information may be determined by the ordinal position in which the information was received.

1. **Anchoring and adjustment, or updating**: The subject may anchor on early information, adjust inadequately for later information. Tversky and Kahneman (1974) demonstrated the common use of this strategy in a task where information was presented simultaneously and subjects anchored on what they read first. Einhorn and Hogarth (1986), Lopes (1982), Lovie (1985), McClelland, Schulze, and Coursey (1986), and Robinson and Hastie (1985) have found that anchoring and adjustment is a useful explanation for how people make diagnostic inferences. In the present task, information is presented sequentially. Two strategies are possible:

   a. **Anchor on initial information:** Subject would remember and anchor on the first piece of information given, and adjust it, taking account of all subsequent information, each time new information is received.

   b. **Updating. Anchor on most recent answer:** Subject would remember and anchor on the previous answer and adjust it taking account of the currently presented information (see Gettys, Kelly, and Peterson, 1973, for a similar theory applied to the Bookbag problem; see Einhorn and Hogarth, 1985). (Note that the result of using this strategy will not necessarily be a universal weighting.)

2. **Primacy:** The first information presented will receive the greatest weight, and less for each subsequent piece of information. This would be expected, for example, with a strategy of anchoring and *insufficient* adjustment. Peterson and Ducharme (1967) found primacy in the Bookbag problem, with the sequence of pieces of evidence $E_1,...,E_n$.

3. **Recency:** Most weight will be put on the most recent information. Pitz and Geller (1970) found such a pattern with the Bookbag problem.

It should be noted that some studies have found no effect of the order of information presentation (e.g., Ricchiute, 1985).

ii. **Content-dependent information weighting patterns.** The subject may use a weighted averaging process, assigning weight according to the kind of information.

1. Most weight on **evidence**.

2. Most weight on **base rate**.

3. Most weight on **reliability**.

4. **Response mode compatibility.** The subject's weighting of the information in the problem may depend on the response mode (see Lichtenstein and Slovic, 1971). Thus, Wyer (1976) proposed the following model:

$$p(H/E) = k_1{}^*p(H) + k_2{}^*p(E/H) - k_3{}^*p(E/\sim H)$$

where $k_2 > k_3 > k_1$. In judging p(H/E), the base rate would get the least weight and p(E/H) would be weighted most because it is most similar to p(H/E) (see also Nisbett, Krantz, Jepson, and Kunda, 1983).

f. **Other composition principles.** The complex arithmetic operations of Krantz and Tversky (1971; see Hypothesis 1-c above) were proposed as possible psychological composition principles, more complicated than averaging or multiplying organizing principles.

2. **Heuristic strategies**. People adopt strategies that simplify the probabilistic inference problem, yet embody a limited appreciation of the meaning of probability. The motivation for the simplification is that it is too difficult to use the probabilities in the normatively correct manner (Tversky and Kahneman, 1974).
   a. **Simplification by the universal neglect of base rate**. Subjects are capable of using most probabilistic information, but they simplify by ignoring base rate information. In particular, they treat the problem as if the base rate = .5, by the principle of insufficient reason.
   b. **Simplification by the neglect of selected information**: People are capable of using any probabilistic information, if there is not so much of it that it overwhelms them. In those situations where there is more information than they know what to do with, they selectively ignore some of it.
      i. **Selective ignoring of base rate.** In those situations where there is base rate information and unreliable evidence, subjects simplify the problem by ignoring the base rate.
      ii. **Selective ignoring of other pieces of information.** Subjects may, for example, ignore the evidence (answering with the prior probability), or ignore the unreliability of the evidence (hence, accepting the evidence fully).
      iii. **Confusion between p(E/H) and p(H/E).** When asked for p(H/E), subjects may give p(E/H), thinking it is exactly the appropriate answer. Eddy (1982, p 254) suggests this explanation: "... the erring physicians usually report that they assumed that the probability of cancer given that the patient has a positive X-ray [P(ca/pos)] was approximately equal to the probability of a positive X-ray in a patient with cancer [P(pos/ca)]. The latter probability is the one measured in clinical research programs and is very familiar, but it is the former probability that is needed for clinical decision making. It seems that many if not most physicians confuse the two." Wyer (1976) and Dawes (1986) provide additional examples.
   c. **Selection of strategies from a repertoire.** In Hypotheses 2-a and 2-b, simplification is viewed as the neglect of selected information, either universally or in particular information overload contexts. Another view is that simplification is produced through the selection of specific strategies in specific situations; "neglect" is simply a side effect. By this view, people have a collection of heuristic strategies, and their selection of a heuristic to use in a situation is guided by an understanding of the meaning of the situation (see Bursztajn and Hamm, 1982; Christensen-Szalanski, 1978, 1980). This hypothesis would say, for example, that people understand the meaning of the probabilistic information in the word problem, and choose relevant heuristic strategies that often produce correct answers, though they are not identical with the normative probabilistic treatment of the problem.

3. **Normative probabilistic reasoning.** The processes people use map onto the processes of probability theory.
   a. **Subjective probabilities.** People reason in accord with the normative probability model; any inaccuracy in their answers is due to their consistent use of subjective probabilities, as in Subjective Expected Utility Theory (Edwards, 1961) or Prospect Theory (Kahneman and Tversky, 1979). That is, they perform the kind of manipulations required by the normative model, except that their conceptions of the probabilities may be slightly distorted.
   b. **Subjective probabilities combined with principled rejection of base rate information.** People reason in accord with the normative probability theory, as above, with the exception that when given base rate information they may consider it to be irrelevant. Specifically, they do not use the base rate information as a prior in the application of Bayes' Theorem. The rejection of the base rate information is based on normative principles. For example, with the Cab problem, people may consider that in

a court one is innocent until proven guilty, and hence statistical evidence about the relative frequency of Blue cabs in the city is not pertinent to the probability of a Blue cab's involvement in the accident (Cohen, 1981); the appropriate prior probability might then be p(H) = .5, in which case the Bayes' Theorem answer for p(H/E) equals p(E/H) (Niiniluoto, 1981).

3. Methods.

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

Three of the seven problems were probabilistic inference word problems, occupying positions 3, 5, and 7 in the questionnaire. The four filler problems required the estimation or calculation of numerical quantities. Most of the problems presented several paragraphs of information, and asked questions after each paragraph. Subjects were instructed to cover each page with a sheet of paper and slide it down to expose only one paragraph at a time. They were explicitly permitted to refer back to earlier paragraphs at any time.

The probabilistic inference word problems are the Cab problem used in previous research (Table 3), the Doctor problem (estimate the probability that a patient has a particular disease; Table 4), and the Twins problem (estimate the probability that a particular twin boy broke a lamp; Table 5). Each problem was divided into four paragraphs, containing the introduction, the base rate information p(H), the evidence, and the reliability of the evidence p(E/H) and p(~E/~H). The subject was asked for the probability of the hypothesis (e.g., that the cab involved in the accident was a Blue cab) and its complement (that it was a Green cab) after each paragraph (see Table 3). The base rate (b), evidence (e), and reliability (r) information within each problem was presented in each of the six possible orders for different subjects. The Cab, Doctor, and Twins problems were also presented in all possible orders, and the information order of the first problem was crossed with the problem order, to create 36 different versions of the questionnaire. The information orders of the second and third problems were linked with that of the first problem, following a pseudo random design which assured that they were different from the information order of the first problem, and that every information order occurred equally often in each ordinal position.

****
Insert Tables 3, 4, and 5 about here.
****

The numerical information in each problem was constant over all questionaires (Cab problem: base rate = .15, reliability = .80; Doctor problem: b = .25, r = .90; Twins problem: b = .20, r = .60). In all problems, the evidence supported the unlikely hypothesis (Cab: the Blue cab; Doctor: the toxic uremia disease; Twins: Stephen), and the reliability of evidence was the same for both hypotheses, p(E/H) = p(~E/~H).

Subjects reported the number of semesters of college math and statistics they had taken, and rated their experience with the content of the seven problems on 1 - 9 scales.

4. Results.

Subjects took from 9 to 54 minutes (mean = 20) to solve the seven word problems. The results of only the three probabilistic inference problems will be reported here. Non-numerical responses were coded as missing. The results pertaining to the subjects' accuracy will be presented first, followed by the evaluation of the list of hypotheses and then incidental findings.

4.1. Accuracy.

The mean answers for each step of each information presentation order are given in Table 6. The accuracy of the subjects' answers at each step can be measured using the mean absolute deviation between their answer and the normatively correct answers. The correct answers (if determinable) for each step of each presentation order are presented in Table 7. The mean absolute deviations between subjects' answer and correct answer are in Table 8.

****
Insert Tables 6, 7, and 8 about here.
****

Before any numerical information was presented to them, subjects usually gave the ".5" answer that the Principle of Insufficient Reason would prescribe. Comparison of Table 6 with Table 7 shows that subjects' answers moved in the appropriate direction at each step of every sequence of information. Consider for example those subjects who received information in the *ber* order for the Cab problem. When given only the base rate (.15 of cabs in city are Blue), their mean estimate of the probability that the cab in the accident was Blue shifted from .50 to .38 (median .30), moving toward the correct answer of .15 (Table 8). The second piece of information these subjects received was the evidence that the witness identified the cab as Blue. Any probability between .15 and 1.0 would now be correct, depending on the subject's confidence in the witness' report. The mean answer shifted from .38 to .79 (median .90). The final piece of information for this group of Ss was the reliability, that the probability is .80 that the witness would say "Blue" (or "Green") given the cab was truly Blue (or Green). The correct answer here is .41, and subjects' mean answer shifted from .79 to .69 (median .80).

The pattern is similar for the remaining Cab problem information presentation orders and for the Doctor and Twins problems. Although the mean answers move in the direction of the correct answers, the shift is too small (suggesting a process of anchoring and insufficient adjustment). As more information is given, the answers are more variable and their means deviate increasingly from the correct answers (Table 8).

Further evidence that subjects answer correctly when given little quantitative information, but become more inaccurate as the information accumulates, is provided by counting the number of subjects who gave the right answers (Table 9). In the *ber* condition of the Cab problem, for example, 42 of 44 subjects gave the best answer (".50") when they had no information, 18 gave ".15" when they had one piece of information, but none gave ".41" with three pieces of information. (With two pieces of information, all 44 subjects were within the .15 to 1.0 range of possible correct answers.) The pattern over all six information presentation orders is that with each additional piece of information to take into account, fewer subjects give the correct answer. The biggest decrease occurs with the third piece of information, where the norm is the complicated Bayes' Theorem. Additional results in Tables 6 to 9 will be discussed below when pertinent to the hypotheses.

****
Insert Table 9 about here.
****

4.2. Evaluation of the hypotheses.

We now evaluate the hypotheses listed above in the light of the present data. The normative probabilistic reasoning hypotheses will be considered first, then the non-normative information integration hypotheses, and finally the heuristic strategies hypotheses.

4.3. Hypotheses involving variants of normative probabilistic reasoning.

Hypothesis 3 in the above list holds that people's reasoning on probabilistic inference word problems follows the normative probabilistic procedures or varies them in a minor way. The analysis of Hypothesis 1-c, in Section 4.4.2 below, includes counts of subjects who applied Bayes' Theorem

either to the correct variables or to their complements. This happened very rarely.

If Hypothesis 3-a is true, a subject's only deviation from the normative probabilistic inference processes would be due to his or her consistent use of subjective probabilities that differ from the objective probabilities. Inspection of the answers when the subject has received only the **b** information (orders **ber** and **bre**, after 1 piece of information, in Tables 1 and 3) shows that the median subject uses exactly the correct number in 5 of the six **b** conditions. Yet these subjects' mean and median answers at the final step (after 3 pieces of information) are substantially different from the Bayes' Theorem answer. In general, subjects are too accurate when given 0, 1, or 2 pieces of information, and too inaccurate with 3 pieces, for there to be a single subjective transformation of the probabilities that they apply consistently in producing all their answers.

Hypothesis 3-b suggests that people consider the base rate information irrelevant in principle, but otherwise follow the rules of probabilistic inference. If they apply Bayes' Theorem when they have all 3 pieces of information, using a prior of .5 instead of the base rate, their normatively correct answer would be identical to the reliability. In this study, as in previous ones, the answer after three pieces of information was frequently the reliability (Cab problem: 39.5%; Doctor problem: 59.8%; Twins problem: 9.2%). The following findings are consistent with the hypothesis that subjects answer correctly when there is no base rate information: with no information, over 85% of all subjects correctly answer ".5"; with **r** only, over 75% of subjects correctly answer ".5"; with r and e, 75% (Cab problem), 85% (Doctor), and 45% (Twins) of subjects correctly give the reliability. However, Hypothesis 3-b predicts that when given the base rate information alone, subjects would ignore it and use the .5 of the Principle of Insufficient Reason. The **b** conditions disprove this, for the base rate was given as the answer by 48% (Cab), 89% (Doctor), and 78% (Twins) of the subjects (Table 8). Hence there is not a general, principled ignoring of the base rate (substituting .5 instead) that could account for the high use of the reliability when subjects have all three pieces of information.

The present results confirm earlier findings that subjects' responses to Bayes' Theorem word problems are not produced using the normative procedures nor their minor variants. However, the results also show that when the situation is simpler because not all of the information pertinent to Bayes' Theorem is present, the modal response is to give exactly the most appropriate answer. This is true when the appropriate answer is the base rate, counter to Hypothesis 3-b.

4.4. Hypotheses involving the non-normative integration of the available information.

Hypothesis 1 holds that people answer word problems by "combining" the available numerical information in some form. Although the correct answers are also produced by combining information, the attention here is on a broad set of possible forms of combination. Each of these hypotheses assumes that the integration is governed by the same rules at each step of the word problem, contingent only on the availability of information.

4.4.1. The use of available numbers.

Hypothesis 1-a holds that subjects use the simplest form of integration -- responding with one number that is available in the word problem and ignoring the others. To test this we will determine how many subjects answered with one of the available numbers. We define a number as "available" if it is present as a number in the text of the problem, present in verbal form (needing to be translated to a number), or available after the operation of complementation. The reliability p(E/H) and the relative frequency of the complementary event p(~H) are given explicitly in the word problem. Some other numbers are implicitly available: the base rate of the target event p(H), which is produced by taking the complement of p(~H), i.e., 1 - p(~H); the prior probability .5; the 1.0 of complete belief in the evidence in favor of H; and the 0.0 of complete doubt. Other available numbers are the answers the subject gave on the first step ("original answer") and on the most recent step ("previous answer") of the problem. Finally, subjects may find an answer by taking the complement of the reliability or of their most recent answer.

Tables 10, 11, and 12 show the number of subjects who used each of these available answers after receiving the first, second, and third piece of information, respectively. For simplicity, if there is no match of a category, the row representing that category is excluded from the tables. Table 13 summarizes by comparing the use of available and non-available numbers. If an answer fit in more than one of these categories, as when .8 is both the reliability and the subject's previous answer, then it is counted only in the first of the categories listed in the table. The order in which strategy categories appear in the table reflects the researcher's assumptions about the subjects' propensities to use the strategies. The danger of this assumption is that when two strategies produce the same answer, if a subject uses the "less likely" strategy, it will be counted as an instance of the "more likely" strategy. While our analysis may inadvertently ignore a strategy on one problem due to such an overlap in the results of two strategies, this neglect is not likely to occur with all three word problems, because with the different numbers used in the problems, the pattern of overlaps among the numbers resulting from applying strategies is different.

****

Insert Tables 10, 11, 12, and 13 about here.

****

After the first piece of information (Table 10), 74.3% (Cab problem), 82.2% (Doctor), and 78.2% (Twins) of the answers made use of one of the available numbers (Table 13). (Note that when the base rate information **b** was given, neither reliability (nor its complement) nor evidence were available, so if subjects used one of these numbers they produced it with a strategy other than the use of an available number. These responses were excluded from the above percentages; and analogously for the **e** and **r** conditions.) Although choice *among* the available numbers is not addressed by Hypothesis 1-a (see Hypothesis 1-c), note that the most frequently used available numbers were the appropriate answers (see Table 7: b for the **b** condition, .5 for the **r** condition, and available numbers in the .5 to 1.0 range for the **e** condition.) But the subjects who did not give these best answers chose one of the other available numbers as the answer as often as they chose non-available numbers.

After the second piece of information (Table 11), the proportions of subjects using available numbers were: Cab: 68.4%; Doctor: 83.3%, and Twins: 76.7% (Table 13). For the **br, er, rb,** and **re** conditions there is a correct answer and, just as occurred after one piece of information, the most frequent response in each of these conditions, for every problem, was this correct answer. For the **be** and **eb** conditions, any answer between b and 1.0 could be correct (see Table 7), and 40% of the subects used available numbers from within this range. Many of the incorrect answers also used available numbers (see Table 13).

After the third piece of information (Table 12), 61.8% (Cabs), 73.2% (Doctor), and 75.9% (Twins) of the subjects used one of the available numbers (Table 13). On the Cab and Doctor problems, the reliability (which is not the correct answer) was used most frequently, and there was substantial use of other available numbers. On the Twins problem, there was less use of the reliability .60 and more of the base rate .20 and the prior probability .50. (We will discuss problem differences in the Incidental Results section, below.)

These results indicate strong support for Hypothesis 1-a. The majority of subjects answer using numbers that are available to the careful reader of the word problem. However, the choice of one available number rather than another is not explained by this hypothesis.

4.4.2. Application of arithmetic operations to available numbers.

Hypothesis 1-a does not account for from 25% to 40% of the subjects' responses. Presumably these answers are produced using more than just one piece of information. This integration may be done through the application of arithmetic operations to the available numbers (Hypotheses 1-b and 1-c) or through a more intuitive, holistic judgment process (Hypotheses 1-d, 1-e, and 1-f).

How well does the set of possible arithmetic operations, applied to the available numbers, account for the answers of those subjects who did not use available numbers? An automated

strategy identification procedure was used to answer this question. It takes advantage of the fact that the inputs are numbers and the strategies are well learned arithmetic operations (presumably executed without error), to identify the strategies without the laborious and unreliable coding process that is typical of verbal protocol analysis. If the analysis is inappropriate because its assumption is not met -- a subject does not apply operations to the available numbers -- the procedure would fail to categorize the subject's answers. The procedure has a weakness, however -- if there are multiple strategies that produce the same answer, it can not determine which strategy the subject used. To check on the relevance of this weakness, we considered the possibility that the subjects might have applied arithmetic operations that happened to produce answers identical to the available numbers.

Table 14 shows the full set of arithmetic operations that were explored. The operations are categorized as: addition, subtraction, multiplication, division, distributive combination (A+P)*U (Krantz and Tversky, 1971), dual-distributive combination A*P+U, and applications of Bayes' Theorem. In addition, the possibility that subjects use conventional probabilities (i.e., thirds, quarters, fifths, tenths, and the extreme probabilities) was considered (Hypothesis 1-d). Every possible application of each arithmetic operation to the available numbers was made, with the constraint that the result be in the 0-1 range, and that the last answer be combined with only the most recently presented information. The "original answer" was excluded since it was ".5" (which *was* included) in over 95% of all cases. As with the analysis of the use of available numbers (Tables 10 to 13), the assignment of answers to these arithmetic operation categories depended on their order in the category list, so that if an answer matched one of the early categories in the list, it would not have a chance to match a later category.

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

In the matching procedure used in this analysis, the first four categories capture the use of available numbers as in the previous analysis. The next set of categories represent the arithmetic operations. The final categories are the conventional probabilities and numbers that fall in the ranges between the points produced by applying the operations to the available numbers. Note that if a number had not yet been presented to the subject, it would not be used in this analysis in producing the set of possible arithmetic operation results.

The numbers of subjects who gave an answer falling in each category after the first, second, and third pieces of information are shown in Tables 15, 16, and 17, respectively. To save space, if for a given problem no answers matched one of the arithmetic operation strategies, e.g., Bayes' theorem or (A+P)*U, then the operation was excluded from the table. Nonetheless, the reader should remember that all the answers in Tables 15, 16, and 17 were tested for possible matches to every one of these categories, in the order specified in Table 14. Table 18 summarizes these results by collapsing the first three categories into the "use of available number" category, and all the categories from Addition through Bayes' Theorem into the "arithmetic operation" category. It can be seen that the popularity of the use of the available numbers decreases slightly as more numbers become available. The unambiguously identified use of conventional probabilities decreases with additional information, while the use of answers that can not be categorized (though rare) increases.

****
Insert Tables 15, 16, 17, and 18 about here.
****

4.4.3. Ambiguity of strategy identification.

In the present analysis, the only data are the subjects' answers; these are assigned to strategy categories because they match the answer that one would get by using the strategy. There is ambiguity in these strategy identifications, because sometimes more than one strategy may produce the same answer. If a subject produces the answer by doing a complicated calculation, the

credit would be given incorrectly to a simpler process -- the use of an available number -- because that comes earlier in the matching sequence. The same ambiguity occurs with the conventional probabilities, the last class of specific answers in the matching process. If the subject uses conventional probabilities in an intuitive judgment process, his or her answer might be counted as an instance of the use of an available number or as the application of an arithmetic operation to available numbers.

The extent of such ambiguity in the data was estimated using the answers to the Cab problem, after three pieces of information (Table 19). Two categorization procedures are compared:

1. A condensed version of the categorization scheme of Table 17, in which the categories are collapsed into (a) the use of available numbers, (b) the use of each particular arithmetic combination of the available numbers, or of conventional probabilities (row labels in Table 19), and (c) "other", including uncategorizable answers as well as all other particular arithmetic combinations, whether they appeared before or after the category of interest in the sequence of matches.
2. A categorization scheme which, for each strategy in turn (arithmetic operation or use of conventional probability), places it first in the sequence of matches.

****
Insert Table 19 about here.
****

Ambiguity between arithmetic operations and use of available numbers. Subjects' answers were identified as using the arithmetic operation of addition of available numbers (first two rows of Table 19) only 22 times (8.6%) when addition was matched following the available numbers, but they were identified as using addition 43 times (16.8%) when it was considered first in the matching sequence. Thus 8.2% of the subjects' answers to this problem are ambiguous with respect to the addition operation. The subject could have produced these answers either by using one of the available numbers (including his or her previous answer), or by adding together two of the available numbers. Although the ambiguity is resolvable in principle, as by studying thinking aloud protocols, it is unresolvable in the present study except by considering which strategy is more plausible. The more likely it is that the subject answered with an available number, rather than answering by adding two available numbers together, the closer our estimate of the percent of people using addition should be to 9% rather than 16%.

Applying this analysis to the other arithmetic operations reveals that the only operation with a large ambiguous component is subtraction. For fully 59% of the answers, it is ambiguous whether the subjects used available numbers or took the differences between available numbers. The subtractions in question are e - c(r) = r [i.e., evidence (1.0) minus the complement of reliability], e - c(b) = b, e - b = c(b), and e - c(.5) = .5. It is reasonable to assume that the subjects use the available numbers (the right hand sides of the above equations), rather than a subtraction strategy (the left hand sides), and so the best estimate of the proportion of subjects who subtracted would be 7% rather than 66% (Table 19).

Ambiguity between the use of conventional probabilities versus the use of available numbers or arithmetic operations. When subjects' answers are assigned to the conventional probability category *after* the categories representing the use of available numbers and arithmetic operations on the available numbers have been matched, only 10 (3.9%) are matched. When the answers are matched to this category first, 189 (73.8%) match the conventional probabilities. It is probably common for subjects to use conventional probabilities to express an intuitive impression, the result of an information integration process, or the result of an anchoring and adjustment process. Hence we can not eliminate the hypothesis that subjects respond with conventional probabilities on the 70% of trials where strategy identification is ambiguous. Therefore, the same data that support the idea that subjects use the available numbers could also be interpreted as supporting their use of conventional probabilities. This ambiguity is resolvable in principle -- unconventional probabilities could be presented in the word problems so that the results of using

available numbers would be in a different class from the rounded results of judgment processes. However, it can not be unequivocally resolved in the present study.

In conclusion, there is much more support for Hypothesis 1-a, which holds that subjects answer using just one of the available numbers in the word problem, than for Hypothesis 1-b, that they combine the numbers using simple arithmetic operations. This conclusion holds even if we credit all ambiguous answers to the more complicated arithmetic operations. However, the results are also consistent with the notion that subjects may be using conventional probabilities (cf. Edwards, 1953) to express a holistic intuitive judgment or the outcome of a process of integrating more than one number. Note also that both hypotheses, using available numbers and using conventional probabilities, are incomplete for they do not explain why the subject picks one number rather than another.

4.5. Weighted average of information.

Hypothesis 1-e holds that the subject's answer may be produced through a weighted averaging process. The "averaging" concept means that all the available information is taken into account, to some degree. The "weight" concept specifies how much relative impact the different input information has, and it is generally taken as a measure of the amount of attention paid to the information, or its importance (Shanteau, 1980).

Weighted averaging has been represented in a variety of mathematical descriptive models (see Hammond, McClelland, and Mumpower, 1980, for a review). There is controversy about whether the averaging process that is embodied in the mathematical models is an accurate reflection of the psychological process by which the judgments are produced, or on the other hand is merely a description of the input/output relations. Hoffman (1960) expressed the position that the model can tell us much of psychological interest about judgments made in an environment, e.g., the accuracy of cue utilization, the relative weights or attention paid to cues, even if it does not describe the psychological process (see also discussion by Einhorn, Kleinmuntz, and Kleinmuntz, 1979). Hammond, on the other hand, argues that for intuitive cognition, at least, the process is indeed some form of mental averaging (Hammond, 1980; Hammond, Hamm, Grassia, and Pearson, in press; Hamm, in press; see also Smolensky, 1986).

The verbal protocol in Table 20 is an example of the kind of process to which the weighted average model might be applied. The subject is first given a .25 base rate, and uses it as her answer. The evidence (which if fully believed would lead to a probability of "1.0") causes her to adjust the answer to .40. Given the reliability next, .70, she moves the answer to .60. This final answer can be considered an average among the inputs. Although a protocol contains information about the justifications offered for the answers following each piece of information, the weighted average approach typically considers only the relation between the input and the output. Further, it does not assume that the integration process heeds the norms of probability. (However, Anderson and Shanteau (1970) and Shanteau (1975) argue for a form of information integration, in the evaluation of gambles, that potentially conforms with those rules.) The weighted averaging process is distinct from the selection of one specific available number (Hypothesis 1-a) and from the combination of information through precise arithmetic operations (Hypotheses 1-b and 1-c). However, a weighted averaging process is compatible with the use of conventional probabilities (Hypothesis 1-d); the average is likely to be "rounded" to such a number (see Table 20).

****

Insert Table 20 about here.

****

To evaluate Hypothesis 1-e, we must ask whether the process the subjects use in answering p(H) on these word problems can be described as a weighted averaging process in either of two senses -- as an anchoring and adjusting or updating process, or as a simultaneous information integration. We lack the data required for the usual methods of testing the weighted average hypothesis by modeling judgments and evaluating the fit of the model: judgments about a large set

of word problems whose dimensions of input information vary systematically. However, some features of the weighted average models can be tested against our data. These features are:

1. Integration. More than one piece of information is taken into account.
2. Betweenness. As the response scale and the information scales are identical, a "weighted average process" must produce an answer that is between the pieces of input information.
3. Nearness. It follows from the "betweenness" assumptions that the relative weight the subject places on a piece of input information will be reflected in the answer's nearness to that input's value.
4. Universality of weights. The subject uses the same integration process, with the same relative weights, when different subsets of the problem information are present. Hence, the ratio of the weights on two pieces of information will be constant, no matter what other information is present. Another way to think of this is that there will be no interactions between types of information, a finding that has been generally true in past research (see Hammond, McClelland, and Mumpower, 1980), with both novice and expert judges (but see Ceci and Liker, 1986, for an exception).

We will test these features of the "weighted average model", as applied to probabilistic inference, with our data.

4.5.1. The integration condition.

The first testable condition is that more than one piece of information be taken into account. In the previous section, it was shown that a large proportion of subjects' answers used numbers that are available in the word problems. Although this seems inconsistent with a weighted averaging model, it does not necessarily eliminate the weighted average hypothesis from consideration. Intuitive integration processes, whether they involve simultaneous consideration of several pieces of information, or a sequence of answers and adjustments, are vague and approximate; hence it is quite possible that a subject will round off the answer provided by a weighted averaging process to the nearest conventional probability. The conventional number to which the intuitive answer is rounded may happen to be available in the problem. Consider the case in which one of two available numbers is given much more weight than the other. The rounding process may lead the subject to respond with this number itself. Next consider the case in which the subject has received all three pieces of information (**b**, **e**, and **r**). Many subjects give the reliability **r** as the answer; yet this number lies between **b** and **e**, and hence it might be the rounded result of an averaging process. Hamm (1987) suggests a response selection mechanism in which rounding to "available" numbers is more likely than rounding to "conventional probabilities". Admitting these considerations, most subjects' answers are consistent with the condition that they use more than one piece of information.

4.5.2. The betweenness condition.

If a subject is using a weighted averaging process, then the answer should be within the range of available numbers. For example, if a baserate of .15 and evidence (p = 1.0) are available, if the subject puts all weight on the base rate information the answer will be .15; if all weight is on the evidence, the answer will be 1.0; any other weighting scheme will produce an answer in between (unless rounded back to an endpoint). This is true when we consider the implicit .5, as well.

To test whether the subjects' answers are consistent with this condition, we assume that the available numbers at a given step in the word problem are the implicit .5 prior probability, plus those pieces of information that have been explicitly presented: **e** (= 1.0), **b** and/or **r** (which have different values for each problem). Only if the subject uses a strategy other than a weighted average could the answer be outside the range of available numbers. The data are largely consistent with this condition. For example, for the no-information condition, the range of available numbers is just the implicit .5. Any weighted averaging strategy would have to produce an answer of .5. For the **b** condition, the range is from the base rate (.15 for the Cab problem) to .5. For the **br** condition, the

Cab problem range is .15 to .80; and for **bre** it is .15 to 1.0. Table 21 presents the number of subjects whose answers were below, in, and above the range of available numbers, for each step of each problem. Most of the answers (between 88% and 99%) fall in the range of available numbers, at every step. Analogous analyses by Fischhoff and Bar-Hillel (1984) and Lichtenstein and MacGregor (1984) showed similar results. While this is very consistent with the weighted average hypothesis, it should be noted that the correct answers, the available numbers, and many conventional probabilities also fall within these ranges. This finding therefore does not eliminate hypotheses of alternative psychological processes.

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

4.5.3. The nearness condition.

The idea that when a subject heavily weights a piece of probability information, the answer will be near to it on the probability scale, follows from the assumption that people understand that the input information and the response are on the same scale. Both the nearness condition and the universality of weights condition must be met for our data to be consistent with the weighted average hypothesis. Hence we can not rigorously test one without assuming the other. We will assume nearness and test universality.

4.5.4. The universality of weights condition.

This condition requires that subjects consistently use the same weighting scheme over different combinations of available types (dimensions) of information (cf. Bar Hillel's (1980) discussion of the integration theory approach). To test this, we must identify possible patterns of weighting the available information, and test whether the data support the hypothesis that any of them are applied consistently, i.e., that the same relative weights are applied to any two kinds of information in different contexts. Subjects might categorize and weight the input information according to the order in which it is received (Hypothesis 1-e-i), or according to its content (Hypothesis 1-e-ii).

4.5.5. Tests of position dependent information weighting patterns.

If Hypothesis 1-e-i is true, then we should find a universal pattern of applying weights according to the ordinal position of the information. The pattern might involve most weight on the earlier information (primacy), uniform weights, or most weight on the later information (recency).

Some results show such a pattern. Consider subjects' answers when given **e** first and **b** second, as compared with the **be** condition, where the identical information is presented in the reverse order. The mean answer for the **eb** condition in the Cab problem is .79, compared with .67 for the **be** condition (Table 6). The difference is significant ($t(84) = 2.17$, $p = .033$). The analogous comparisons are .93 to .85 for the Doctor problem ($t(82) = 1.71$, $p = .092$) and .52 to .45 for the Twins problem ($t(84) = 1.17$, $p = .245$). Since the e information (1.0) is at the upper extreme of the input scale, this pattern indicates that the answers are nearer to the more recently given information, i.e., that more weight is given to the most recent information.

A second example of weighting according to ordinal position occurs when all three pieces of information are given. The **b** information is always the lowest, **r** intermediate, and **e** highest. Table 6, "3rd info" column, shows that the mean answers are lowest when **b** is given last, again indicating that the ordinal position of the information influences the weight the subjects give it. For the Cab problem, the mean of the **erb** and **reb** conditions is .63 (medians are .68 and .75, respectively) and the mean of the remaining conditions is .69 (median = .80), $t(254) = 1.96$, $p = .051$. For the Doctor problem, when b is last the mean answer is .67 (medians .70 and .75), compared with .81 (median .90) in the other four conditions ($t(257) = 4.78$, $p < .001$). For the Twins problem, the difference between the mean answer when b is presented last (.36) and when it is presented in an earlier position (.43) is also significant ($t(260) = 2.97$, $p = .003$). This heavy weighting of recent information is a factor that has influenced the results of previous studies on the Cab problem, in which it was found that many subjects answer with the reliability, which is too high. The order of information in

those earlier studies was **ber**, in which the reliability is presented last. Changing information presentation order in the word problems would decrease, though not eliminate, the bias of "ignoring base rate". Bar-Hillel's (1980) generalization, "the median and modal responses were consistently based on the indicator alone, demonstrating the robustness of the base-rate fallacy", would need to be modified in detail, if not in spirit.

These two findings are consistent with the notion that people use a weighted averaging approach, weighting the information according to ordinal position. Because more weight is given to recent than to early information, they counter the "insufficient adjustment" part of the anchoring and adjustment or updating versions of the weighted averaging hypothesis: Hypothesis 1-e-i-2 is eliminated.

Other comparisons, however, do not support the notion that people weight the available information only according to the order in which it is received. When the **r** information is given first, following the implicit .5, it is virtually ignored: subjects continue to answer ".5". But when the **b** or **e** information is given first, the subjects move a substantial portion of the distance from .5 to the new information (Table 6, "1st info" column). This indicates that subjects do not universally pay more attention to the most recent information.

The **rb** and **br** conditions support the same conclusion. Table 6, "2nd info" column, shows that the means for these two conditions are nearly identical ($p > .5$ for both the Cab and Doctor problems; $t(83) = 1.82$, $p = .072$ for the Twins problem, where **rb** elicited lower answers (.27) than **br** (.34)), and much nearer to the base rate than to the reliability. This is consistent with a primacy weighting pattern in the **br** condition, and a recency weighting pattern in the **rb** condition. Overall, then, the results are not consistent with the universal use of a weighted average process that is based on the ordinal position in which the information is received. Thus Hypotheses 1-e-i-1-a, 1-e-i-2, and 1-e-i-3 are rejected.

Hypothesis 1-e-i-1-b is a special case because it assumes the subject updates, i.e., integrates the latest information with his or her previous answer rather than with previous information. However, the finding that the answers in conditions **rb** and **br** are almost identical, while the answers in **eb** and **be** are different, is not consistent with this hypothesis.

Therefore we can reject all variants of Hypothesis 1-e-i. To say that people answer probabilistic inference word problems using a weighted averaging process with a sequential information weighting pattern does not adequately account for the data.

4.5.6. Tests of content dependent information weighting patterns.

The alternative possible basis for weighting information is according to the kind of information (Hypothesis 1-e-ii). For example, subjects might give more weight to base rate information than to reliability. If so, their answers in the **br** and **rb** conditions should be nearer to the base rate information, no matter whether it is presented first or second. In fact, the mean answers are nearer to base rate than reliability (Table 6, "2nd info" column, "bre" and "rbe" rows), which supports this hypothesis.

But some results are not consistent with this hypothesis. First, consider the **r**, **br**, and **rb** conditions. Here the reliability information receives very little weight (Table 6) and is seldom used (Tables 10 and 11). But in the **re** and **er** conditions, as well as in all conditions with all three pieces of information (Table 12), the modal answer is the reliability (except for the Twins problem with three pieces of information). Thus the reliability information seems to be weighted differently in different contexts, in violation of the universality of weights condition.

A second implication of the universality condition is that the ordinal relations between weights of pairs of dimensions should be transitive. That is, if the weight on Dimension A is greater than the weight on Dimension B in all situations in which information on the two dimensions is available, and the weight on Dimension B is greater than the weight on Dimension C, then the weight on

Dimension A should be greater than the weight on Dimension C.

In order to test the transitivity of dimension weights, a measure of weight is needed. One can be developed using the "nearness" assumption. The weighted averaging hypothesis holds that a subject's answer on the probability scale is determined by averaging the input information, which is presented on the same scale. We can express the general weighted averaging model as follows:

$$R = \frac{w_0 \times X_0 + w_1 \times X_1 + w_2 \times X_2}{w_0 + w_1 + w_2}$$

where X_i is the input information on dimension i, X_0 is the initial impression, the implicit .5, and w_i is the weight on dimension i. (The inclusion of the initial information makes this a little complex, but we will show later that the result of the analysis is the same whether or not this is included.) The nearness condition says that the answer will be nearer to the information that is given more weight. "Nearness" is the opposite of "distance", and so we can estimate the weights in this model, from a single answer R (or the average of a group of subjects's answers on the same problem), by measuring it as some form of complement of distance. $1 - |R - X_i|$ is such a measure. Therefore the model of how the answer is produced can be expressed as:

$$R = \frac{\sum (1 - |R - X_i|) \times X_i}{\sum 1 - |R - X_i|}$$

and the weight on dimension k can be expressed as:

$$\frac{1 - |R - X_k|}{\sum 1 - |R - X_i|}$$

This expression can be used to measure the relative weights of the evidence, reliability, and base rate dimensions in the conditions where only two of these three pieces of information have been presented. When only the evidence and reliability have been presented (conditions **er** and **re**), the mean answer for the Cab problem is .76 (the mean of the **erb** and **reb** rows, Table 6, "2nd info" column). The estimate from the nearness model for the weight on reliability is .39 (see Table 22). This is derived by substituting R = .76, X_0 = .5, X_e = 1.0, and X_r = .80 into the nearness model, above. The estimate for the weight on the evidence is .31. (The estimate for w_0, the weight on the .5, is .30). Thus $w_r > w_e$.

****
Insert Table 22 about here.
****

When only the reliability and the base rate have been presented (conditions **rb** and **br**), the mean answer for the Cab problem is .31. The weight estimates from the nearness model are w_b = .39 > w_r = .24. Since $w_b > w_r$ and $w_r > w_e$, transitivity of relative weights predicts that w_b will be greater than w_e. However, when only the base rate and the evidence have been presented, the mean answer is .73, and the weight estimates are w_e = .38 > w_b = .22. Table 22 shows that this intransitive pattern holds with several alternative formulations of the nearness model, as applied to the Cab problem results. It holds whether or not the initial impression X_0 is included. It holds when the nearness measure is derived by subtracting the absolute distance from the range of available information, (Inf_{max} - Inf_{min}), rather than from the full probability range of 1. The intransitivity is also found with the Doctor problem, but not with the Twins problem. [One reason for this exception, as we shall discuss in Section 4.8 below, is that people spontaneously know that twin identifications are unreliable, and so the **eb** answers in the Twins problem are not really done "without any **r** information".]

This intransitivity of relative weight is a failure of the universality of weight condition and hence a disproof of the hypothesis that subjects consistently apply a weighted averaging process to the information, according to the type of information. There is no possible universally applied weighted averaging scheme on the reliability, base rate, and evidence information dimensions that could be consistent with the intransitive pairwise weighting that has been observed here. This is true whether or not the implicit .5 information is included in the model. Therefore, Hypotheses 1-e-ii-1 through 1-e-ii-4 are eliminated.

The finding that the universality of weights condition is not met has one of two implications. It may be that subjects do not produce their answers to probabilistic inference word problems by applying a weighted average process to the information. If so, all variants of Hypotheses 1-e should be rejected. On the other hand, subjects may use a weighted averaging process, but the weighting scheme may not be universal, i.e., different relative weights may be applied to the information dimensions in different contexts. Thus the weight placed on the base rate information, relative to the evidence information, by the subject in Table 20 may change when the reliability information is received. If this implication is correct, then a new question arises: what determines the weights in the different situations? Bar-Hillel (1980) suggests that people use (weight) information that they perceive as "relevant". She manipulated this relevance by changing elements within the full word problem. Our results show that if subjects are indeed taking weighted averages, the weights depend on which dimensions of information are present. Therefore the perceived relevance of one kind of information may depend on the presence of other dimensions of information. Another possibility is a response mode effect (see Wyer, 1976). Before evidence information has been given, subjects are asked to respond with p(H); when evidence has been given, their responses are by definition p(H/E). The base rate or prior probability, which has the form p(H), may be given more weight than reliability before the evidence has been presented, when the response mode is p(H), which is similar to it; while the reliability p(E/H) is given more weight than the base rate after the evidence is available, when the response mode is p(H/E).

In searching for what might determine "relevance", we should not neglect the normative theory. Consider the correct answers when one is faced with pairs of pieces of information (see Table 7, "Two Pieces" column). If one has only the base rate and reliability information, the base rate is the correct answer, so using nearness as a measure of weight, $w_b > w_r$. If one has the reliability and the evidence, and assumes a prior probability of .5, the correct answer is the reliability, so $w_r > w_e$. If one has the evidence and the base rate, the correct answer can be anywhere between the base rate and the evidence, depending on one's assumptions about the reliability. In most inference communication situations, the presumption is that the reliability of evidence is fairly high (if it were not, the communicator would be expected to say so; see Kahneman and Tversky, 1982), and so the answer would be closer to the evidence (1.0) than to the base rate. With these reasonable assumptions, $w_e > w_b$. Hence the intransitive pattern observed in our data is the same pattern that would be seen if the subjects were sensitive to normative considerations. Although as we have shown above the subjects are not universally using the normatively correct procedure to produce their answers, they may well be using heuristic strategies that are broadly sensitive to these normative considerations.

4.5.7. Other composition principles.

Hypothesis 1-f is that people combine the available information using organizing principles based on multiplication of the information inputs, or more complicated organizing principles, rather than averaging. There is little support for multiplying, because the answers are rarely below the range of the available numbers (see Table 21), which would characterize the products of numbers between 0 and 1. We will not analyze the more complicated organizing principles here, because our data are not adequate to test them. The high use of available numbers, and the failure of the simpler weighted averaging and multiplying hypotheses to account for the data, make it unlikely that subjects use the more complicated composition principles.

4.6. Simplifying and heuristic strategies.

The final set of hypotheses that we shall evaluate holds that people answer probabilistic inference word problems by using heuristics, "strategies of simplification that reduce the complexity of judgment tasks, to make them tractable for the kind of mind that people happen to have" (Kahneman, Slovic, and Tversky, 1982, p xii). In this view, any way of producing an answer can be considered a "strategy", including selecting an available number, producing a weighted average of inputs, and calculating Bayes' Theorem. A heuristic strategy is relatively simple, and can be executed using humans' limited memory and coordination capacities. In this view, the heuristic strategies of interest in probabilistic inference word problems are less complicated than the mathematical operations of Bayes' Theorem.

The heuristic strategies theory is broad and powerful. The strategy concept includes both intuitive (e.g., the use of weighted averages, or of conventional probabilities) and analytic (e.g., the application of mathematical operations) processes (Hammond, Hamm, Grassia, and Pearson, in press). The strategy used in a given situation may be newly invented, explicitly selected from a repertoire (a decision process; see Christensen-Szalanski, 1978; 1980; Bursztajn and Hamm, 1982), or automatically applied (a quasi-perceptual process; see Tversky and Kahneman, 1982). Any number of strategies might be applied to a given situation; different people may apply different strategies; different people may apply the same strategies for different reasons; a person may apply different strategies in different situations, or in the same situation at different times. Consequently it is very difficult to disprove the general heuristic strategies theory, though specific candidates may be rejected. However, some common features distinguish any heuristic strategy from the processes considered under Hypothesis 1.

1. Contingency. Different strategies may be adopted in different situations, depending on aspects of the task situation. The implication for probabilistic inference word problems is that there is no expectation that the same process will be used when different combinations of information are available.

2. Accuracy. The very definition of heuristic strategy implies that over situations (if not in each situation) accuracy will be better than random, yet less than perfect. This holds whether the mechanisms for strategy invention and selection involve conscious justification (Slovic, 1975) or trial and error (March, 1978). This implies that answers will be more accurate than what would be expected if any one of the strategies were to be applied in all situations.

3. Variation in accuracy. The accuracy of the outcome will depend on the task and the strategy. The details of the task situation will determine the accuracy of the application of the chosen strategy to the task. It should be noted that the features of task that influence the adoption of strategy may or may not be related to the features of task that determine its accuracy (Rose, 1974). As a consequence, variation in accuracy between situations to which the same strategy is applied may be expected.

To evaluate Hypothesis 2, we will seek evidence for these general features in the subjects' responses. Out of the infinite set of possible strategies, we will focus our analysis on strategies that have been named in the literature and/or that are easily identified in our data, which lack process observations and repeated judgments.

4.6.1. Contingency in the neglect of information.

An easy strategy for simplifying a situation is to ignore some of the available information. For example, people answering probabilistic inference word problems have been characterized as *neglecting* base rate, which means "the base rate is either ignored or grossly underweighted" (Tversky and Kahneman, 1982, p 153). When people use an available number as their answer, they neglect the other information (unless, as discussed in Section 4.5.2 above, they round off to the available number, or average two numbers that are on either side of it). In Section 4.4.1 above we counted the number of people who used available numbers in each situation; here we will focus on the neglected numbers in the same situations, and look for variations in the pattern of neglect across situations.

Subjects in this study used the base rate when it was the only information presented (see discussion of Hypothesis 3-b, Section 4.3 above), and so we can reject Hypothesis 2-a, that they universally neglect the base rate information. However, it is possible that people use the base rate in some situations but neglect it in others. We may test whether the base rate information has an influence in each situation by comparing the subjects' answers with it and without it, using within-subject and between-subject t-tests. Within subjects, the difference between the answers before and after the presentation of the **b** information can be tested (Table 23). Between subjects, the answers of all subjects who have received a set of information without **b** can be compared with the answers of all subjects who received the same information plus **b** (Table 24).

****

Insert Tables 23 and 24 about here.

****

The first row of Table 23 shows that for the Cab problem, subjects with no information answered .50, and then when given base rate information (.15) their mean answer was .34. Therefore, when the base rate is the only information available, it is not neglected. Tversky and Kahneman have found a similar result (unpublished data from 1973, cited by Bar-Hillel, 1980). The second row shows that when given base rate information following evidence, subjects' mean answer decreased significantly from .85 to .67. In fact, for each of the five possible situations in the Cab problem in which subjects received base rate information following a previous answer, the new answer was significantly lower. This is true for all three problems.

Table 24 presents comparisons between subjects. The first row indicates that the mean answer was .76 for subjects who were given evidence first and then reliability, compared with .69 for a different group of subjects who were given baserate, evidence, and then reliability. The difference is marginally significant ($t(85) = 1.94$, $p = .055$). Base rate information had a significant impact in every comparison in the Twins problem.

The pattern for the Cab and Doctor problems is that whenever the base rate was the most recently presented information, the mean answer was significantly lower (reflecting appropriate attention to base rate) than the answer when base rate information was lacking, but if the base rate was presented earlier, the difference is not significant (though $p < .15$ in seven of the 9 comparisons in question, and $p < .30$ in the other two, all in the expected direction). This reflects the recency effect discussed above. In conclusion, subjects do not generally *ignore* the base rate information. But it is used more in some situations than in others. It had a highly statistically significant impact on their answers when it was presented last and when presented along with only the reliability information. However, its impact was not statistically significant in other situations, as revealed by the comparisons **er** versus **ber** and **ebr**, **re** versus **bre** and **rbe**, and **e** versus **be**, for the Cab and Doctor problems. Because recency plays an important role in the use of base rate, one might propose that this is a universally applied strategy that can account for the results. However, it was shown in Section 4.5.5 above that recency does not influence the use of all information equally. A hypothesis that recency governs the use of base rate information, but not of other information, would in itself be a "contingent strategies" hypothesis. The neglect of base rate therefore seems to be produced by strategies that are not applied universally.

Similar analyses were done to determine whether evidence and reliability information have an impact in all situations. Evidence is not at all neglected, with significance levels of $p < .001$ in every comparison (data not included). In some conditions reliability has a large effect, and in other conditions it has no effect (Tables 25 and 26). For all problems, the mean answers for the **er** and **re** conditions are significantly lower than for the **e** condition. This is true in both within subject and between subject comparisons. On the other hand, the reliability information does not affect the answer when it is presented alone (in the **r** condition; with the exception of the Doctor problem, where the difference is due to a very few answers that are less than .5) or if it is presented in conjunction with the base rate information (in the **rb** and **br** conditions; with the exception of the Twins problem; see Section 4.8 below). Reference to Table 7 shows this lack of effect to be correct. Finally, note that when reliability information is presented last (Table 25), the direction of change

depends on whether evidence was presented first and base rate second (the change is a slight increase) or base rate was first and evidence second (the change is a significant decrease). This interaction may be attributed to two occurrences of the recency effect. First, before the reliability is presented, evidence and base rate have been presented in one of two orders. The answers tend to be near to the most recently presented information. It happens that the mean answer in the **eb** condition is below the reliability value (which is yet to be presented), but the mean answer in **be** is near or above it (and hence many individuals' answers are above it). Second, when the reliability is subsequently presented, the answers move toward it: shifting up for the **eb** condition and down for the **be** condition.

****
Insert Tables 25 and 26 about here.
****

In conclusion, there is evidence for variation in the degree to which the subjects' strategies neglect base rate and reliability information in different situations. While the variation in the utilization of base rate may be due to recency effects, the variation in the use of reliability is not, and thus it is unequivocally an instance of the contingency that is characteristic of a heuristic strategies account of people's answers to probabilistic inference word problems.

A second example of contingent strategy use is the utilization of available numbers. We showed in the discussion of Hypothesis 1 that many subjects respond using numbers that are presented in the word problem. Let us now analyze the selection of particular available numbers. Consider the relative rate of use of the base rate and the reliability numbers, when both are available. Table 11 shows that when only reliability and base rate were available (conditions **rb** and **br**), many more subjects used the base rate (Cab problem: 30 used base rate and 2 used reliability; Doctor: 61 to 2; Twins: 46 to 2). Yet when evidence information was also available (Table 12), the selection shifted (Cab problem: 12 subjects used base rate and 101 used reliability; Doctor: 7 to 155; Twins: 52 to 24). Though in both cases the strategy is a version of "select an available number", the tendency to select reliability compared with base rate changed when the evidence information was added. This is a clear instance of contingent strategy use.

4.6.2. Increased accuracy due to contingent strategy use.

In the previous example, the use of different strategies in different situations contributes to subjects' accuracy. When only the base rate and reliability information are available, it is appropriate to respond with the base rate, as most subjects do (see Table 7). When all three pieces of information are available, neither extreme, base rate nor evidence, is the right answer; a subject who recognizes this and is committed to using an available number would do better to select one that is in between these extremes. The reliability and the implicit .5 of the principle of insufficient reason are the alternative candidates. Table 12 shows that the reliability was the modal response after all three pieces of information had been presented in the Cab and Doctor problems, while .5 was the most frequent response in the Twins problem. These responses are relatively near to the correct (Bayes' Theorem) answers for the Doctor problem (.75) and the Twins problem (.27), though not for the Cab problem (.41).

A second example of increased accuracy due to contingent strategy use is the intransitive weights applied to pairs of information dimensions, discussed above (Section 4.5.6). The same intransitive pattern of relative weight was demonstrated to be characteristic of normative thinking about these situations. This demonstrates contingent strategy use, and increased accuracy as a consequence.

4.6.3. Variation in accuracy of the same strategy in different situations.

The third feature of a "heuristic strategies" explanation is that the accuracy of a strategy will vary across situations. A general example may be seen in Tables 8 and 9: answers were less accurate when the problem had more information, i.e., when the situation became more complex. The biggest decrement is between two and three pieces of information, when the norm becomes Bayes'

Theorem. It could be that subjects use strategies that cope quite well when the problem is relatively simple, but are inadequate when it becomes more complicated. Thus, despite the observation in the previous section that strategies vary when there are two pieces of information and the answer is consequently more accurate, still the general pattern seems to be that the strategies do not change enough in response to task changes.

A specific example of a heuristic strategy which works in some situations, not in others, is the strategy of interpreting reliability p(E/H) to mean the same thing as p(H/E) (Hypothesis 2-b-iii). The hypothesis that errors in probabilistic inference are produced by such a confusion has been put forth as an explanation of the poor performance on probabilistic inference word problems when all three pieces of information are present (Eddy, 1982; Dawes, 1986; Wyer, 1976). The common use of reliability as a response when all three pieces of information are present (see Table 12) can be attributed to the use of this heuristic strategy -- people think that the p(E/H) information is exactly what the question is asking them for. Note, however, that reliability is also the most common response when only the evidence and reliability information are available (Table 11; conditions **er** and **re**). We noted in Tables 8 and 9, above, that these answers are right. However, this may be by accident. It just happens that in the **er** and **re** conditions, p(H/E), calculated with Bayes' Theorem assuming a prior probability of .5, yields an answer equal to the reliability, p(E/H) (see Niiniluoto, 1981), and so the heuristic of using p(E/H) for p(H/E) produces an exactly correct answer. [This would not be the case if sensitivity, p(H/E), were different from specificity, p(~H/~E), in these problems.] The conditions for applying this heuristic are that both evidence and reliability information be present. Without the evidence E, reliability as p(E/H) would not be perceived as pertinent (see the **no** information, **r**, and **(br)** conditions), and the required response would be perceived as p(H) rather than as p(H/E), and so no confusion would be possible.

This last example is notable because the same heuristic is used when there are two pieces of information, as when there are three, and it is equally "confused" or "inappropriate" in both cases. However, the answer happens to be right when there are two pieces of information, with no credit due to the subject's understanding, while it is very wrong (in the Cab and Twins problems) when there are three pieces of information.

It will be helpful at this stage to adopt a more formal representation of a contingent strategies explanation. We shall model it as a production system in the style of an OPS5 program (Brownston, Farrell, Kant, and Martin, 1985), in which subjects' behavior in various situations is controlled by a collection of rules or "productions" which are used when appropriate. Each rule consists of two parts, conditions and actions. The conditions are compared with the situation, as represented in working memory. If the conditions in a rule match the situation, then the rule's actions are taken. For example, one production for the present case is:

If
  I have been asked to answer p(H)
  I have base rate information
Then
  Answer p(H) = base rate.

This rule would be selected in cases where base rate information is present, and result in the base rate being given as the answer.

What would happen if the subject had both base rate and evidence information? In addition to the first production, another production, such as

If
  I have been asked to answer p(H)
  I have base rate information
  I have evidence
Then
  Answer p(H) = 1.0

would match this situation. When two rules match, which one should take action? This is decided by

a process called "conflict resolution", in which one of the matching rules is selected for application, on the basis of conflict resolution principles. One of these principles is that the production whose conditions are most specific is selected. Hence, the second production would be used, and the answer would be 1.0.

Table 27 shows the conditions and actions of three sets of productions. The first set represents normative behavior. It will produce the correct answer in every situation of our study. The second set of productions represents the typical subject on the Cab and Doctor problems, while the third set captures the typical answer on the Twins problem. These will produce the most popular answers in every condition. Note that rules A, B, and H are the same in all three sets. For rules C and E, the conditions are the same but the actions are different. The special rules for the typical subject on the Cab and Doctor problems are called C', D', and E'. The typical Twins problem subject uses C", D", E', and F". Rules F and G appear only in the normative set. Thus, the production system representing normative reasoning in probabilistic inference word problems has rules A through H; the production system representing the typical Cab and Doctor problem subject has rules A, B, C', D', E', and H, and the typical Twins problem subject production system has rules A, B., C", D", E', F", and H[1]. Other rule sets could be used to model nontypical subjects.

****

Insert Table 27 about here.

****

Consider first the set of normative strategies. Rule A provides for a response of estimating baserate information if one is asked for p(H) but has no pertinent information. This rule will match every situation; however, only when there is no baserate b nor evidence e information available will it be selected (Table 28), due to the specificity principle. The actual estimating is done by rule H, which produces a value for b -- .5. The conditions for Rule B are then met, and it produces an answer equal to b, which is .5. Rule B is used when there is baserate information, without evidence and with or without reliability. Rule C will fire (be matched and then selected) when only evidence information is available. Its action is to call for estimates of both reliability and baserate information. (Note that the production system would still produce the right answer if only one of these, say "estimate r" were called for.) In the present system, the only thing that can be done in response to a request for an estimate of base rate or reliability is to make a subjective judgment. This is embodied in productions G and H, which "guess" about reliability and baserate, respectively. A reasonable judgment about baserate or prior probability, in the absence of specific base rate information and prior knowledge, is the .5 of the principle of insufficient reason. A guess about reliability will depend on what the subject believes about the reliability typical of evidence in such situations. There is no normatively prescribed answer for this, and therefore the specification in the set of normative productions must allow for a judgment based on the subject's own knowledge about the content area of the word problem.

****

Insert Table 28 about here.

****

Rule D is applied when there is information about the baserate and evidence, and Rule E is applied when there is information about reliability and evidence. Finally, with information about all three pieces of information, rule F is executed. Its action produces an answer by calculating Bayes' Theorem. This normative set of productions embodies a contingent strategies theory, because different actions are done in different situations. However, because the actions are the best possible according to our normative standards, we would not call this a model of a heuristic strategies theory; there is no deviation from the best response.

[1]This suggests a multidimensional definition of "simplicity" in cognition: not only can one rule be simpler than another, but a small set of rules can be considered simpler than a large set.

The second and third sets of productions, representing the contingent strategy use of the typical subject on each problem (see Column 1 of Table 6, Column 1 of Table 9, and Tables 10, 11, and 12), can be considered to embody a heuristic strategies theory because their answers are not correct in all situations. Rules A, B, and H are the same as for the normative production set. Rules C' and C" will substitute for rule C, rules D' and D" for D, and rule E' will substitute for rules E and F in the Cab and Doctor problem production systems, and rules E' and F' will substitute for them in the Twins problem production system. Rule C' provides that the response 1.0, indicating complete acceptance of the evidence, will occur whenever the evidence is given, unless the reliability is also given, in which case Rule E' applies and the response will be r.

Note that the only estimates made by the typical Cab and Doctor problem subject are of the prior probability when neither the b nor the e information has been given (in the Null and r conditions). In contrast, the typical Twins problem subject estimates b in the e condition, and the normative strategy additionally involves estimates of r in the e and (be) conditions, and estimates of b in the e and (er) conditions. This estimating or searching behavior is something that can be looked for in the verbal protocols or process traces of experts who presumably use better strategies than novices. This theory specifies the conditions under which such behavior can be expected.

The production system formalization of the typical subject's strategies (actions) and the rules which cause them to be used in the various situations in this study (conditions) allows us to address in a specific manner a general issue concerning people's performance on probabilistic inference word problems. The issue is whether the poor performance on probabilistic inference word problems is due to a process in which base rate information is neglected (e.g., Tversky and Kahneman, 1982; Bar-Hillel, 1980) or to a process in which reliability information is misunderstood (e.g., Dawes, 1986; Eddy, 1982; Wyer, 1976). The issue is important because of its implications for how to aid people to make better inferences. For instance, Fischhoff and Bar-Hillel (1984) have explored methods of calling attention to the base rate information, which make more sense in the context of a process of neglecting base rate than in one of misunderstanding reliability.

The above production system model, which exactly produces the typical subject's responses, attributes the error to a dual process. The driving factor is that the reliability is misunderstood. Specifically, the subject thinks it to be the information that he or she is asked to produce. The second factor is that, in the context of the misinterpretation of the reliability information as the target information, the base rate information no longer seems relevant, for the answer is already in hand. in this account, the neglect of base rate information happens not because of an erroneous process of assessing the relevance of information, a process which wrongly decides that the bae rate is less relevant than p(E/H). Rather, the neglect occurs because of a reasonable assessment that the base rate information is less relevant than the "p(H/E)". The error is in a confused interpretation of the p(E/H) reliabiilty information as p(H/E).

The typical subject's response in the (ber) and (er) conditions of the Cab and Doctor problems, guided by rule E', is the reliability, which the subject thinks is the appropriate answer because of a confusion between p(E/H) and p(H/E). This misunderstanding is general. That is, there are no rules in which the typical subject has the opportunity to confuse the reliability information but does not do so. Although some of the rules in the production system are compatible with a correct understanding of the reliability information, none of the rules requires that the subject distinguish successfully between the two conditionals. Rule B, applied in the (rb) conditions, would produce the same (correct) answer if the subject confuses p(E/H) with p(H/E). Because there is no evidence information E, neither p(E/H) nor p(H/E) would be relevant here. Similarly, Rule A, applied in the r condition, would produce the correct answer whether or not the reliability information is confused.

The reliability concept plays a special role in conditions e and (eb), for the subjects are not given reliability information here. Normatively, their answers can be anywhere between the 1.0 of complete acceptance of the evidence and neglect of the base rate, and the .5 or b of complete acceptance of the base rate and neglect of the evidence. Their exact answer depends on the reliability that they attribute to the evidence. No reliability information is given; no mention of reliability or unreliability has been made (though 2/3 of the subjects had just answered another word

problem which used reliability concepts, and the subjects bring to the Twins problem the knowledge that identification of twins is unreliable). What reliabilities do the subjects spontaneously assume? This can be estimated by turning Bayes's Theorem inside out, that is, by solving for it (Table 29). For example, the mean answer for p(H/E) in the Cabs problem, be condition, is .79. We can solve for p(E/H)[2], yielding .96 as an estimate of the reliability assumed by the mean subject. This is not to say that the subjects are inverting Bayes' Theorem in their heads; that is implausible given that they can not apply it straight. But the estimates produced by this procedure have two valid uses. First, whether the subjects know it or not, these are the reliabilities that they are assuming; if they do not think these are right then they should adopt another strategy for answering the p(H/E) question. Second, meaningful comparisons can be made between the estimated reliabilities for different problems. The modal responses for the Cab and Doctor problems in Table 29 indicates that in the absence of specific reliability information, the subject assumes the evidence is completely reliable. On the other hand, in the Twins problem subjects assume that the evidence is fairly unreliable. In the absence of specific reliability information, the modal estimate is that the evidence is completely unreliable. This comparison between the Cab and Doctor problems, in which subjects assume high reliability in the absence of specific reliability information, and the Twins problem in which low reliability is assumed, serves as a bases for explaining problem differences (Section 4.8.1 below).

****

Insert Table 29 about here.

****

In conclusion, the production system representing the typical subject's response is consistent with two important processes a fundamental misunderstanding of reliability information every time it is presented, and a neglect of base rate information when it is present in the context of the misunderstood reliability information. Further, when the reliability information is not presented, most subjects in the Cab and Doctor problems acted as if they did not think about the possibility that the evidence might be unreliable. (In the Twins problem, however, they did think of reliability when it had not yet been presented.) Nowhere in this production system is there a rule that embodies a misunderstanding of the presented base rate or a lack of appreciation of the concept of base rate when specific information about it was lacking. Rather, the neglect of base rate information is produced by the firing of rules which embody the misunderstanding of the reliability information. Given the misreading of p(E/H) as p(H/E), the neglect of base rate may well be appropriate. This interpretation is consistent with Tversky and Kahneman's (1982) and Bar-Hillel's (1980) general account, which holds that people do not know how to integrate statistical (base rate) information with single case (evidence) information. Even if some or all of these subjects understand the different meanings of p(E/H) and p(H/E), they do not know how to integrate all their information using Bayes' Theorem. This account differs from Tversky, Kahneman, and Bar-Hillel's account in the importance accorded to the confusion of p(E/H) and p(H/E).

A previous study by Christensen-Szalanski and Beach (1982) is consistent with this theory. Student subjects were given "experience" by being exposed to 100 cards, each representing a medical case that either had or did not have a disease, before being given a word problem analogous to those used here. In one condition the cards had information only about whether the patients had the disease, and hence they conveyed base rate information. This experience made no difference in the students' accuracy on the word problem -- they still neglected the base rate and

---

[2]Let p = p(H/E), b = p(H), r = p(E/H), and p(E/~H) = 1 - p(E/H). Bayes' Theorem is

$$p = \frac{b \times r}{b \times r + (1-b) \times (1-r)}$$

Solve for r as follows:

$$2pbr + p - pb - pr = br$$

$$2pbr - br - pr = pb - p$$

$$r = \frac{pb - p}{2pb - b - p}$$

used reliability as their answer. In another condition, the case information included both the true disease and the test result (positive or negative). Thus the subjects experienced not only the base rate but also the reliability. The accuracy of the students in this condition was substantially improved by this experience, in part because fewer of them mistakenly used the reliability as the response. It was therefore the experience of the reliability, and not the experience of the base rate, that enabled these subjects to answer the probabilistic inference word problem more accurately. Similarly, the most successful attempt to train people to do better on probabilistic inference word problems has involved display of the 2 by 2 table relating evidence (E and ~E) to hypotheses (H and ~H), which expresses both p(E/H) and p(H/E) (Lichtenstein and MacGregor, 1984).

How can the theory that errors in probabilistic inference are due mainly to confusion in the interpretation of reliability information be reconciled with previous findings that changes in the distribution of answers can be produced (and the rate of neglect of base rate decreased) by manipulating the perceived relevance of the base rate information (Bar-Hillel, 1980; Fischhoff and Bar-Hillel, 1984)? I acknowledge that not all subjects use heuristic strategies as represented in the typical subject production systems. Other subjects may use weighted averaging strategies in which weights on the dimensions of information can be influenced by manipulations of perceived relevance. And even within the heuristic strategies approach, some subjects may not confuse p(h/E) with p(E/H). Their strategies for combining the statistical and case information may well be responsive to factors that show the base rate to be relevant.

4.6.4. Conclusions concerning heuristic strategies.

This study has found strong support for Hypothesis 2-c, which holds that subjects have a number of different strategies and select different ones in different situations. The heuristic strategies explanation captures some general features of the behavior of the average subject: the use of different ways of combining information in different situations; the fact that these changes tend to parallel the changes demanded by normative considerations; and the variation in accuracy when the same strategy is followed in different situations.

In addition, there is support for Hypothesis 2-b-iii, the "confusion hypothesis", that subjects interpret the reliability p(E/H) as if it were p(H/E). This not only explains the modal answers in the conditions where all three pieces of information are given, it is also consistent with their behavior when less information is present. Paradoxically, the most appropriate use of the reliability concept seems to have occurred when no specific reliability information was presented. Here subjects used their general knowledge concerning the reliability of evidence and they did not deviate far from the range of correct answers, though they were perhaps a little optimistic about the quality of the evidence in the Cab and Doctor problems, and a little pessimistic in the Twins problem.

4.7. Summary of the tests of the hypotheses.

We have shown that Hypothesis 3, that people's answers on probabilistic inference word problems are produced by processes that are variants of Bayes' Theorem, simply does not account for the data. Hypotheses 1-a, 1-b, 1-c, and 1-d together describe a large proportion of the answers; however, the most successful of these, Hypotheses 1-a (use of an available number) and 1-d (use of a conventional probability), are very broad, and hence are also consistent with weighted averaging (Hypothesis 1-e) and with the use of heuristic strategies (Hypothesis 2). A weakness of Hypotheses 1-a and 1-d is that they do not offer a basis for predicting *which* available number or conventional probability will be selected. Weighted average and heuristic strategies explanations can do so.

In evaluating the weighted average hypothesis (1-e), we discovered that there was no possible pattern of weights that could account for the results in the conditions where only two dimensions of information were presented. It would therefore be necessary to speak of different patterns of weights being used in different situations. This contingency is one of the characteristics of the heuristic strategies explanation (Hypothesis 2), which is a more general hypothesis than weighted averaging, for it covers many forms of strategy (including the selection of available numbers, and

weighted averaging). Overall, the heuristic strategies account (Hypothesis 2) explains the data more successfully than Hypotheses 1 or 3. As embodied in a production system representing the typical subject, the heuristic strategies theory exactly predicted the modal subject's selection of available numbers, while the weighted averaging explanation, in order to be compatible with the modal answers, would have had to resort excessively to the notion of "rounding off". Further, when the heuristic strategies approach was modeled explicitly it strongly suggested that the primary source of errors in probabilistic inference word problems is in the subjects' misunderstanding of reliability information. It also makes predictions about situations in which search may be expected from expert subjects. In conclusion, the heuristic strategies approach offers both the most accurate explanation of the results and the account that provides the largest number of new insights into the phenomenon and the most ideas for future research.

4.8. Incidental Results.

This section presents results pertaining to issues distinct from the evaluation of the hypotheses in Section 2.2: differences between the problems, subject stability between problems, subject factors that may influence accuracy, and the accuracy of the subject's most likely guess.

4.8.1. Differences between problems.

It has been noted several times above that the results on the Twins problem differ from those on the Cab and Doctor problems. This may be attributed to two causes. First, due to an error in producing the problems, a number of the Twins problems had sequences of paragraphs that did not present a coherent narrative. Second, subjects have prior knowledge about the unreliability of identification of identical twins.

The effects of the incoherent narrative. 221 subjects received versions of the Twins problem in which the narrative was coherent when the paragraphs were presented in the order baserate, evidence, reliability, but when presented in different orders there were references in early paragraphs to events that were not specifically mentioned until later paragraphs. Although these events are all well understood within a "babysitter" frame, this incoherence might affect subjects' responses. The questionnaire was corrected, and 41 subjects (11 in the **reb** information presentation order, and 6 in each of the other orders; all with the Twins problem presented first) were given the coherent version. The mean responses of the two groups are presented in Table 30. The answers for three of the six information presentation order conditions, following three pieces of information, are different by an amount greater than or equal to .10 (**bre**, **erb**, and **rbe**). T-tests of the differences between incoherent and coherent versions, for all conditions, are presented in Table 31. A few of the differences were significant. Several people with the coherent version gave answers less than .5 in the **r** condition; in the **br** and **bre** conditions (which involve the same subjects), the mean answers were lower for the subjects with the coherent version. For this reason, all the above analyses relied primarily on the results of the Cab and Doctor problems. A further concern is whether following the incoherent version of the Twins problem affected the answers on the other problems. Comparisons of the means were made and no significant differences were found.

****

Insert Tables 30 and 31 about here.

****

Problem differences due to prior knowledge of the reliability. The three problems differ in the extent to which the typical subject knows that the evidence is unreliable. Everyone knows that it is hard to tell twins apart. On the other hand, most people blindly trust medical technology and do not spontaneously wonder about the reliability of medical tests. The knowledge of unreliability of evidence in the Cabs problem can be expected to be intermediate (see the evidence paragraphs of the Twins, Cab, and Doctor problems in Tables 5, 3, and 4). This prior knowledge of unreliability can be expected to make a difference in the responses in those conditions where evidence is presented without reliability information. Table 32 shows the effect of getting the evidence

information under these conditions; **e** is compared with the no-information condition, and **be** is compared with the **b** condition, for each problem. The data for the incoherent and coherent versions of the Twins problem are shown separately. The effect of the evidence when that information is presented first in the Doctor problem is a difference of .47, the mean answer shifting from .49 to .96. The analogous effect is .35 in the Cab problem, and .15 in the Twins problem. This agrees perfectly with the prediction that there would be least expectation of evidence unreliability in the Doctor problem, and most in the Twins problem. The pattern is identical when evidence is given following base rate information. There seems to be little effect of the incoherent versus coherent version of the Twins problem in this.

****
Insert Table 32 about here.
****

General comparisons. Inspection of Tables 10, 11, and 12 shows that there are major differences in the distribution of answers over available numbers, between the Twins problem and the other two. For example, Table 13 shows that 40% of the Cab problem answers and 60% of the Doctor problem answers after three pieces of information use reliability as a response, while only 9% of the Twins problem answers do. Table 12 shows that subjects use the base rate and .5 instead. Although there is a significant effect of the incoherent version of the Twins problem, it is relatively minor and is not sufficient to account for these differences. The prior knowledge of the unreliability of identifications of twins seems to be the cause. Note that not only does this knowledge influence the use of evidence before reliability information is presented (Table 32), but it also seems to make people distrust the evidence *and the base rate* throughout the problem.

4.8.2. Subject stability over problems.

The attempt to identify subjects' strategies is based on the assumption that subjects use the same strategies on different problems. Without such stability of strategy, little prediction would be possible. (However, see Hamm (1987) for a theory of probabilistic answer selection or strategy use.) Therefore it is of interest whether subjects in this study meet this assumption -- do they use the same strategy on different problems?

Because of the counterbalancing of orders of information presentation, it is not meaningful to test the stability of subjects' strategies after one or two pieces of information. Only after three pieces of information will the subjects be solving the same problem. Four analyses will be made: correlations of answers between problems, correlations of accuracy between problems, and seeking evidence of the use of a strategy that produces the same class of answer on pairs of problems, defining "class of answer" at two levels of detail.

Correlations of answers between problems. If a subject consistently uses strategies that produce high or low answers, relative to the answers of other subjects, then there should be a positive correlation between subjects' answers on different problems. The correlations for the answers after three pieces of information are as follows:

	Doctor	Twin
Cab	.07	.04
Doctor		-.03

These small and nonsignificant correlations provide little evidence for consistent use of strategies that produce high or low answers.

Correlations of accuracy between problems. If a subject consistently uses a strategy that produces a relatively accurate or inaccurate answer, then there should be correlations between the accuracy index (the absolute value of the deviation from the correct answer) between problems. The correlations between the accuracies of the answers after three pieces of information are as follows:

	Cab	Doc

Doc	-.01	
Twn	-.01	-.02

Evidently there is little consistency in the accuracy of subject's responses. A separate index of accuracy involves determining whether the subject's choice of hypothesis would be right if forced to choose (see Section --- below). We assume that subjects would choose the hypothesis favored, i.e., if p(H) were greater than .5, choose H, and if p(H) were less than .5, choose ~H, with random choice if p(H) were equal to .5. The answers after three pieces of information can be rescaled in this way. For the Cab problem and the Twins problem, the correct answer is ~H; for the Doctor problem, the correct answer is H. Hypothetical forced choice answers are scored as 1 if correct, 0 if incorrect, and .5 if the original answer was .5 and forced choice would have been random. Correlations between these estimates of accuracy, for the three problems, are:

	Cab	Doc
Doc	-.06	
Twn	.11*	-.03

There is a significant, by tiny, relation between the accuracy of their imputed guesses on the Cab and Twins problems. As a general conclusion, analysis of accuracy reveals little evidence for subjects' stability of strategy use across problems.

Answers in the same answer class: detailed classification scheme. The next analyses assign subjects' answers to a classification scheme and test whether the subject's answer falls into the same category on two different problems. An example of a category is "the use of the reliability number from the problem". Presumably, if the subject uses the reliability number on two problems, it would be due to using the same strategy on both problems. This assumption is weaker when the category is a range, such as "numbers greater than the reliability but less than 1.0".

The first categorization scheme (top of Tables 33, 34, and 35) is complex. It used more than 16 categories, but only those on which at least one subject used the category on at least one of the two problems are listed in the tables. The first category is the use of the base rate number as the response. One subject used base rate on both the Cab and the Doctor problems. The expectation is calculated using the row and column marginals (divided by the total) in a 16 by 16 table, which presents the number of subjects who used each possible combination of categories on the two problems. The named categories in Tables 33 to 35 are the diagonals from such tables. Categories are either exact numbers (such as the complement of the base rate, the last answer, or the prior, which is the subject's answer when no information had been presented, usually .5), or ranges of answers (such as all number that are less than both the base rate and the prior). A Chi-squared test of the number of subjects falling on the diagonal in this table is conducted. These are the subjects who used the same category in this scheme. This test shows that a significant number of subjects used the same category on the Cab and Doctor problems, but that there was no such stability between either of these and the Twins problem. Although the test is statistically significant for the Cab and Doctor problems, the 79 subjects who used the same class of answer on both problems is only a small proportion of the 265 subjects, very close to the 60.7 who would be expected to have done so by chance.

****

Insert Tables 33, 34, and 35 about here.

****

Answers in the same answer class: simple classification scheme. The analogous analysis was done with a less complex classification scheme (lower table in Tables 33, 34, and 35). The scheme used only the baserate, .5, the reliability, and 1.0, plus the intervals between them. Note that this will inflate the estimated stability, because people whose answers on two problems fell in the same interval will be counted as using the same strategy, even if different strategies that produce similar answers were used. However, there may be some connection between answers that fall into the same interval. For example, if subjects are using the same strategies but rounding, estimating, or using weighted averaging processes, their stability would be lost in the more detailed

categorization scheme. The results show evidence of strategy stability between the Cab and Doctor problems, and between the Cab and Twins problems, but not between the Doctor and Twins problems.

In conclusion, the evidence is not strong that subjects use the same strategy (i.e., one that produces exactly the same kind of answer) on different problems. Although there is statistically more use of the same category of response than chance, still only a small minority of subjects show this stability.

4.8.3. Subject factors that influence accuracy and strategy choice.

Are there any features of subjects that influence how accurately they respond to probabilistic inference word problems? One candidate is the amount of experience subjects have with the content of the word problem. Subjects were asked to rate their experience with each of the seven problems on the questionnaire. A normalized or relative rating for each subject is constructed by dividing the rating for each problem by the total experience the subject reported for all problems. The correlation between these experience ratings and accuracy (1 - absolute deviation of subject answer from correct answer) after three pieces of information, for each problem, is:

	Score		
	Cabs	Doctor	Twins
Raw			
Experience	-.07	-.03	-.09
p	.118	.320	.073
Relative			
Experience	-.11*	-.04	-.18*
p	.041	.281	.002

For each problem, it was found that the more experience one had had with the content of the problem, the less accurate one's answer was. This is statistically significant for the Cabs and Twins problems, when the relative experience score is used. This result, though surprising, is related to the finding that higher I.Q. subjects neglect the base rate more than lower I.Q. subjects do (Maya Bar-Hillel, personal communication).

Other information we have about subjects includes their year in college, the amount of time they took to complete the total questionnaire (seven problems), which is presumed to reflect the time they spent on each problem, and the number of semesters they have taken of college mathematics and college statistics. The range on the last two was quite low, with the mode at 0 semesters. The correlations between these factors and accuracy of response at each amount of information, for each problem, is shown in Table 36. Only three of the correlations (out of 60) are significant, each concerning the Doctor problem. Because accuracy is measured as 1 - absolute deviation, a negative correlation indicates that the more of the factor, the less accurate the answer. Thus, the students who were taking introductory psychology later than their freshman year, and those who took more time on the questionnaire, gave less accurate answers after one piece of information on the Doctor problem, and those who took longer gave more accurate answers after all three pieces of information on the Doctor problem. Thirty four of the correlations in the table are negative, 26 positive. There seems to be little relation between subject factors and accuracy, in the range of variation of these college student subjects.

****

Insert Table 36 about here.

****

Investigation of whether subject factors influence strategy choice will be carried out when subjects with mathematical training or with experience in the field of the word problem have responded to analogous word problems.

4.8.4. Accuracy of forced choice.

If someone were forced to choose one of the two hypotheses, they would (in the absence of principles of "innocent until proven guilty") probably select the answer that they already favor, that is, they would choose H if p(H) were greater than .5, ~H if p(H) were less than .5, and they would choose randomly if p(H) = .5 (but compare Slovic, 1975). This would be consistent with a principle of going with "the preponderance of the evidence".

Given the subject's answer p(H), it is possible to calculate the probability that the subject's choice would be correct (see Table 37). For example, with the Cab problem, at one piece of information, if the information is the base rate .15, then the probability that "blue" is right is .15. If the person gave an answer of p(H) greater than .5, they would be correct (according the the best information available) only .15 of the time; if less than .5, they would be right .85 of the time; and if equal to .5, they would be right .50 of the time. Those cases where the answer is indeterminate were excluded (e.g., if given base rate and evidence; see Table 7) from the mean. This happened with one out of three conditions when there was one piece of information (e), and with two out of six conditions when there were two pieces of information (be or eb).

****

Insert Table 37 about here.

****

The left half of Table 37 shows that the mean probability of being correct after one piece of information on the Cab problem is .593 (N = 174); the mean probability is .742 after 2 pieces of information, but falls to .447 after 3 pieces. The right half of Table 37 shows the probability of being correct after 3 pieces of information if one had calculated Bayes' Theorem. this is .59 for the Cab problem, compared with a probability of .41 if the subject were "diabolical", that is, trying to be wrong. The mean of the subjects' probabilities, .447, is distressing close to the worst they could do. This is not so for the other problems. For the Doctor problem, where the Bayes' Theorem answer is on the same side of .5 as is the evidence rather than the base rate (because the evidence is of such high reliability), the probability that the average subject would be correct is only 7% lower than the probability that the best subject would be; and it is .43 above the probability of the diabolic, purposefully wrong subject. In the Twins problem, where the Bayes' Theorem answer is on the opposite side of .5 from the evidence, the pattern is similar: the mean subject's probability of being right is .14 below the best subject's, and .32 above the worst subject's.

It is helpful to recognize that there are only a few possible probabilities a subject could have here: .5, the same probability as the ideal subject, and the complement of the ideal subject's. Tale 38 shows the number of subjects whose imputed chances would give them each probability of selecting the correct hypothesis. There are five probabilities in the Two Pieces of Information condition because, although subjects can have only three possible probabilities, the high and low in the two conditions are different.

****

Insert Table 38 about here.

****

When there was only one piece of information it was the reliability for half the subjects, and most of them answered .5. This accounts for the approximately 50% of subjects who had a .5 chance of choosing the correct hypothesis. Besides these (and the others who guessed .5), many more people guessed in the right direction than the wrong direction on each problem (nearly 4 to 1 on the Cab problem, 13 to 1 on the Doctor problem, and 18 to 1 on the twins problem).

When there were two pieces of information, there were two ways subjects might make the worse choice, depending on whether they were in the **(br)** condition or the **(er)** condition. [Remember that **(be)** subjects are excluded from this analysis because the true probability is indeterminate.] In the **(br)** condition, subjects would have the base rate chance of being right if they guessed high (opposite the base rate). In the **(er)** condition, would have the reliability chance of being right if they selected H, and the complement of the reliability chance of being right if they

guessed low (opposite the evidence). With the cab problem, 4.7% sat on the fence and 85.3% guessed right, leaving only 10% to guess wrong. With the doctor problem, 91.5% guessed right, 4.6% sat on the fence, and 3.8% got it wrong. With the twins problem, 71% guessed right, 21% sat on the fence, and 7.9% guessed wrong.

With all three pieces of information, if subjects guessed low (on the Cab and Twins problems) or high (on the Doctor problem) they were on the same side as the Bayes' Theorem answer, and hence they had the high probability of being right. The result on the Cabs problem is notable: 18.8% of subjects would have guessed right, 3.9% would have had a .5 chance of being right, and 77.3% would have made the wrong choice and had the lower probability of being right. This lower probability is only 18% lower than the higher probability available if one guesses right (an example of the principle that the harder the decision, the less it matters). In the Doctor problem, 84.6% of the subjects would have made the better choice if forced. In the Twins problem, 59.2% would have make the better choice.

The Cabs problems seems special, in that its base rate was low enough, and reliability low enough, so that subjects' answers were on the wrong side of .5, and hence they would probably make the wrong choice if forced.

5. Discussion.

The present project has intensively analyzed the responses of college student subjects on probabilistic inference word problems. The understanding we derive here is useful for knowing what kind of performance can be expected from people on these types of problems and how they can be trained or aided to do better. Cohen (1981) argued that people's performance on word problems (concerning the effect of sample size on variability of result) is of little import because naive subjects can not be expected to know technical statistical laws. Tversky's (1981) reply is relevant here: "This argument misses a major point about psychological research. Of course, naive subjects are not expected to formulate or prove laws of statistics or geometry. However, the psychologist is very interested in whether naive subjects have learned from lifelong experience that nonrepresentative results are more frequent in small than large samples" (p 355). Similarly, with probabilistic inference word problems we can not expect that naive subjects understand Bayes' Theorem and are able to apply it to the word problem. But it is of interest whether lifelong experience has taught them that statistical information (base rate) can be combined with case information (evidence), whether they know how to adjust a prior degree of belief appropriately given unreliable evidence, and whether they can distinguish the implications of two conditional probabilities, the probability that a particular type of evidence would be observed given that a hypothesis were true, and the probability that a hypothesis would be true given that particular evidence were to be observed.

5.1. Significant findings.

This project studied the probabilistic inference word problems intensively by requiring subjects to respond before and after each of the pertinent pieces of information: the base rate, the evidence, and the reliability of the evidence. Because the information was presented in different orders to different subjects, answers to eight possible combinations of information and 16 possible presentation orders were observed. This enabled the testing of a number of general hypotheses concerning naive subjects' responses and accuracy on these problems. Given previous demonstrations that mathematically sophisticated subjects (see Tversky and Kahneman, 1971) and subjects expert in the content area of the problems (Eddy, 1982) make errors on word problems, models of naive subject behavior may well be very pertinent to expert behavior.

A notable result is that subjects paid attention to base rate information. When it was presented without case evidence, subjects relied on it heavily, as is appropriate. When it was presented along with evidence information, subjects paid attention to it, although on the average they accorded it insufficient weight (speaking non-technically) and many subjects ignored it completely. This supports Tversky and Kahneman's (1982) and Bar-Hillel's (1980) accounts of people's performance

on the full problem as a "neglect" of base rate, yet shows that people do understand the base rate information, and consider it pertinent to the question, contra Cohen (1981).

To a great extent, people responded using numbers that were directly or indirectly available in the presentation of the word problem. Although these available numbers were often appropriate responses for the problem when one or two pieces of information were available, the tendency to respond using available numbers continued into the situation where all three pieces of information were available and the appropriate response was not an available number. In contrast to the frequent response with available numbers, it was fairly infrequent that subjects applied mathematical operations to available numbers to produce their responses.

A result that is of only minor theoretical interest but has practical import for some situations is that subjects gave more weight to the most recently presented information, either when selecting an answer from the numbers available, or in combining the numbers into an answer.

The major theoretical result of this study is its support for a contingent strategies theory, which holds that people apply different response strategies in different situations. The hypothesis that they combine the information using the same weighted averaging policy in all conditions was rejected. If they use intuitive weighted averaging policies, they use different weights when faced with different combinations of types of information. But the frequent use of the available numbers suggests that the response is not produced through intuitive averaging but rather through the application of rules prescribing the selection of available numbers. The contingent strategies expression of this insight holds that different rules for selecting available numbers will be applied in different situations. This theory was modeled as a production system program with six productions, or rules specifying that particular actions be taken in particular situations. This model exactly predicted the most frequent response in every situation (every combination of information) for the Cab and Doctor problems, and a variant with seven rules predicted the modal responses for the Twins problem.

Inspection of the rules in the production system model, and their conditions of application, showed that there was no rule that involved a misunderstanding of the base rate information. Rather, if base rate was ignored it was because other rules had already been applied and produced an answer without considering it. This is consistent with Bar-Hillel's (1980) notion that information is used in producing an answer on these problems when it is "relevant", and that this relevance depends on the situation. Inspection of the rules that tended to dominate, i.e., to be applied before those rules that incorporate the base rate information, showed that they were consistent with a confusion between reliability p(E/H) and the posterior probability p(H/E). That is, these rules would have been more appropriate had their input been p(H/E). Therefore the misunderstanding of the reliability information seems to be an important factor in the neglect of base rate.

On the other hand, even if subjects had understood the reliability correctly, there is no rule in the production system that would have allowed them to produce the correct answer with all three pieces of information. Thus, even if we were to train naive subjects to correctly interpret p(E/H), this would only remove a block to correct performance, by removing a reason for ignoring base rate information; it would not provide a method for correct performance, or even approximately correct performance. Ongoing study of the strategies that mathematics experts and substantive experts apply to probabilistic inference word problems may fill this gap.

These results show that on probabilistic inference word problems, people have difficulty applying their lifelong experience so that they can integrate statistical (base rate) with case (evidence) information. The confusion between the two conditionals, p(H/E) and p(E/H), as expressed in the problems, seems to be an important cause of this difficulty.

Another contributing factor may be the conventions of communication. People expect to be given the information they need to solve puzzles. Experience in educational contexts that use word problems has taught us that we will be given problems that we are expected to be able to solve using the information at hand and principles that we have recently been taught; if no principles have been taught, then all the information that is needed can be found in the problem (Fischhoff and

Bar-Hillel, 1984). And there is no penalty for guessing. A further aspect of the education game is that teachers may expect students to fail on a new type of problem, to provide motivation for the next lesson. These considerations suggest that the tendency to say that the probability a hypothesis is true is one of the numbers available in the situation may occur only in probabilistic inference word problems, not in real world situations that require probabilistic inference. In my ongoing and planned work, the word problems are being presented in ways designed to break up this hypothesized mental set:

1. Subjects are asked to think aloud about the problem.
2. Subjects are given probabilistic information in one code (numerical probabilities or verbal probabilities) and asked to respond using the other code, to prevent the easy use of a response available in the word problem.
3. Subjects are given the opportunity to select the information that they would like to have, and to explain the reasons for their selections.

5.2. Implications of the neglect of base rates in word problems.

It has been shown that people do not successfully integrate base rate probabilities and probabilistic descriptions of the reliability of the available evidence when solving probabilistic inference word problems. How should this affect our understanding of the prospects for improving human rationality through the use of probabilities as measures of the degree of belief in propositions? Four positions may be identified: the results may be irrelevant, we should improve the way that information is presented in the situation, we should help people make better inferences, or we should train people to increase their statistical understanding.

The results are irrelevant. This position holds that, because word problems are not realistic, they tell us nothing significant about human rationality. People may well be rational in their lives yet fail to get the right answers on word problems (Christensen-Szalanski and Bushyhead, 1981; Cohen, 1981). By this argument, the research on probabilistic inference word problems has nothing to tell us about improving human rationality.

It is incorrect to argue that the findings of word problem research on probabilistic inference are irrelevant. There are some real world situations very analogous to these word problems (e.g., when a physician is consulted about another physician's patient, he or she spends a few minutes listening to a verbal description; Dawes, 1986); further, the same neglect of base rate observed in the Cab problem has been seen in medical textbooks (Eddy, 1982) and in physicians' answers to word problems (Cascells, Shoenberger, and Grayboys, 1978).

A second reason that the demonstrated neglect of base rate may be considered irrelevant has to do with our finding that subjects are "on the right side of .5", that is, if the subjects were forced to choose which hypothesis to bet on, they would usually pick the one which is more probable, by the normative calculations. Even in the Cab problem, where people would choose the less likely hypothesis, the difference in the probability of being right (.41 if choose wrong, .59 if choose right) is small. However, there are a number of situations in this world in which the exact probability is important because one must do more than just "pick the best bet". One disease may be more serious than another, and the choice of therapy may depend on probability estimates. The automated radar defense system on the U.S.S. Stark in the Persian Gulf was programmed to translate evidence into action in a way that was insensitive to the prior probability of an attack. to prevent embarassing false alarms, the system was routinely shut off when no attack was expected. And it was shut off when the ship was hit by an Iraqi missile. Whether or not the responsible officer's judgment that no attack was expected was correct, it is clear that the system would work better in practice if it were capable of taking judged prior probabilities into account, and if the people who use it understood the base rate concept and could use it properly. It is also clear that the responsible officer's court martial and trial by press would be more fair if the judges and the public understood the concepts of base rate and reliability of evidence well enough to apply it to this "word problem".

Focus on the situation. In this view, if we know the factors that make the base rate seem relevant to someone, we can construct important inference situations in such a way that available information about base rate will be used appropriately.

However, focussing only on situations is not practical. We do not control the presentation of information in all important probabilistic inference situations. Were the principles of relevance widely known, it is as likely that situations would be constructed to mislead people (as can occur, for example, in the selling of insurance) as to help them reason correctly.

Focus on tools and aids. The advocates of decision aiding hold that it is not realistic to expect people to be able to do complicated mathematics, such as Bayes' Theorem, in their heads, if they lack training or intellectual tools such as formulas and calculators (von Winterfeldt and Edwards, 1986). The author recently sat in on an undergraduate statistics lecture in which Bayes' Theorem was taught, and tried to solve two problems in his head before the lecturer worked them through. While he succeeded (within .02), the intensity of the required mental effort surprised him. In a word problem where people are not given the opportunity to look up the formula or use calculators, the implicit message is "we do not really expect you to get this right"; even statisticians get the message and fail on the Cab problem (see Bar-Hillel, 1980). By this view, the way to assure human rationality would be to make the intellectual tools available and train people to use them, or to have decision analysts available to help them, just as in historical eras when most people were illiterate scribes could be hired to write for people.

Although it may be possible to aid those who repeatedly are confronted with socially important probabilistic inference problems, by providing them with (a) appropriate intellectual tools that they are trained to use, (b) decision analysts, or (c) decision support systems, this is not a practical universal solution to the irrational mental strategies revealed by the word problems.

Focus on people's statistical understanding. By this view, people who are either trained with statistics, or expert in the area of the word problem, tend to use more statistical concepts, more accurately, in answering probabilistic inference word problems (Nisbett, Krantz, Jepson, and Kunda, 1983). Rather than changing the situation to increase the perceived relevance of the information that should be used, statistical training could increase the subject's ability to discern the relevent information (see Meehl and Rosen, 1955; Widiger et al, 1984); mere "mechanical manipulations" are not enough (Fischhoff and Bar-Hillel, 1984). Lichtenstein and MacGregor (1984) have explored various methods for training people to do well on probabilistic inference word problems, and found that those methods that both provide intellectual tools (such as the 2 by 2 table of evidence any hypothesis possibilities) and explain why these work, helped best.

Enhancing people's understanding of the pertinent statistical concepts and how they apply in the particular situations seems to be the most general approach to improving people's ability to be rational through the use of probability measures of degree of belief. Its effects would be felt in many situations, not only in those where the presentation of information was specially controlled, or the problem solver had special tools or helpers available.

In order to improve intuitive reasoning, it is helpful to describe unaided reasoning, for this is the base upon which improvements must be laid. The present work has identified the role of the confusion between (E/H) and p(H/E) as a block to successful integration of statistical and case information, and has shown that people use strategies contingent on the types of information presented. These findings will be useful to the project of improving probabilistic inferences.

6. Bibliography.

Anderson, N.H. (1981). Foundations of Information Integration Theory. New York: Academic Press.

Anderson, N.H., and Shanteau, J. (1970). Information integration in risky decision making. Journal of Experimental Psychology, 84, 441-451.

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

Birnbaum, M.H., and Mellers, B.A. (1978). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. Paper presented at the Midwestern Psychological Association, 1978. [cited in Birnbaum, M.H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. American Journal of Psychology, 96, 85-94.]

Borgida, E., and Brekke, N. (1981). The base rate fallacy in attribution and prediction. In J.H. Harvey, W.J. Ickes, and R.F. Kidd (Eds.), New Directions in Attribution Research, (Vol. 3). Hillsdale, N.J.: Erlbaum.

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

Bursztajn, H., and Hamm, R.M. (1982). The clinical utility of utility assessment. Medical Decision Making, 2, 161-165.

Cascells, W., Schoenberger, A., and Grayboys, T. (1978). Interpretation by physicians of clinical laboratory results. New England Journal of Medicine, 299, 999-1000.

Christensen-Szalanski, J.J.J. (1978). Problem solving strategies: A selection mechanisms, some implications, and some data. Organizational Behavior and Human Performance, 22, 307-323.

Christensen-Szalanski, J.J.J. (1980). A further examination of the selection of problem-solving strategies: The effects of deadlines and analytic aptitudes. Organizational Behavior and Human Performance, 25, 107-122.

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

Christensen-Szalanski, J.J.J., and Bushyhead, J.B. (1981). Physicians' use of probabilistic information in a real clinical setting. Journal of Experimental Psychology: Human Perception and Performance, 7, 928-935.

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

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

Edwards, W. (1953). Probability preferences in gambling. American Journal of Psychology, 66, 349-364.

Edwards, W. (1961). Behavioral decision theory. Annual Review of Psychology, 12, 473-498.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.),

Formal Representation of Human Judgment. New York: Wiley, 1968, pp 17-52.

Einhorn, H.J. (1985). A model of the conjunction fallacy. Unpublished manuscript, Center for Decision Research, Graduate School of Business, University of Chicago.

Einhorn, H.J., and Hogarth, R.M. (1985). A contrast/surprise model for updating beliefs. Graduate School of Business, University of Chicago.

Einhorn, H.J., and Hogarth, R.M. (1986). Decision making under ambiguity. Journal of Business, 59,, Number 4, Part 2, S225-S250.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology: Human Perception and Performance, 1, 288-299.

Fischhoff, B., and Bar-Hillel, M. (1984). Focusing 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.

Gettys, Charles F., Kelly, Clinton, III, and Peterson, Cameron R. (1973). The best-guess hypothesis in multi-stage inference. Organizational Behavior and Human Performance, 10, 364-373.

Hamm, Robert M. (in press). Clinical intuition and clinical analysis: Expertise and the cognitive continuum. In Jack Dowie and Arthur Elstein (Eds.), Professional Judgment. Cambridge: Cambridge University Press.

Hamm, Robert M. (1987). A model of answer choice on probabilistic inference word problems. Paper presented to Mathematical Psychology Society meetings, Berkeley, California, August 6-8.

Hammerton, M. (1973). A case of radical probability estimation. Journal of Experimental Psychology, 101, 252-254.

Hammond, K.R., Hamm, R.M., Grassia, J.L., and Pearson, T. (in press). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. IEEE Transactions on Systems, Man, and Cybernetics.

Kahneman, D., and Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3, 430-454.

Kahneman, D., and Tversky, A. (1972). On prediction and judgment. Oregon Research Institute Research Monograph, 12(4). Cited in Fischhoff and Bar-Hillel, 1984.

Kahneman, D., and Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80, 237-251.

Kahneman, D., and Tversky, A. (1982). On the study of statistical intuitions. Cognition, 11, 123-141.

Krantz, D.H., and Tversky, A. (1971). Conjoint-measurement analysis of composition rules in psychology. Psychological Review, 78, 151-169.

Lichtenstein, S., and Fischhoff, B. (1980). Training for calibration. Organizational Behavior and Human Performance, 26, 149-171.

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

Locksley, A., and Stangor, C. (1984). Why versus how often: Causal reasoning and the incidence of judgmental bias. Journal of Experimental Social Psychology, 20, 470-483.

Lopes, L.L. (1982). Toward a procedural theory of judgment. (Report WHIPP 17). Madison: University of Wisconsin.

Lovie, P. (1985). A note on an unexpected anchoring bias in intuitive statistical inference. Cognition, 21, 69-72.

March, J. (1978). Bounded rationality, ambiguity, and the engineering of choice. Bell Journal of Economics, 9, 587-608.

Marks, D.F., and Clarkson, J.K. (1972). An explanation of conservatism in the bookbag-and-pokerchips situation. Acta Psychologica, 36, 145-160.

McClelland, Gary H., Schulze, William D., and Coursey, Don L. (1986). Valuing risk: A comparison of expected utility with models from cognitive psychology. Unpublished paper, University of Colorado.

Meehl, P.E., and Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. Psychological Bulletin, 52, 194-216.

Niiniluoto, I. (1981). L. J. Cohen versus Bayesianism. The 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.

Peterson, C.R., and DuCharme, W.M. (1967). The primacy effect in subjective probability revision. Journal of Experimental Psychology, 73, 61-65.

Peterson, C.R., and Miller, A.J. (1965). Sensitivity of subjective probability revision. Journal of Experimental Psychology, 70, 117-121.

Pitz, G.F., and Geller, E.S. (1970). Revision of opinion and decision times in an information-seeking task. Journal of Experimental Psychology, 83, 400-405.

Ricchiute, D.N. (1985). Presentation mode, task importance, and cue order in experimental research on expert judges. Journal of Applied Psychology, 70, 367-373.

Robinson, L. B., and Hastie, R. (1985). Revision of beliefs when a hypothesis is eliminated from consideration. Journal of Experimental Psychology: Human Perception and Performance, 11, 443-456.

Rose, G. (1974). Unpublished paper, Department of Management Sciences, University of Iowa.

Shanteau, J.C. (1972). Descriptive versus normative models of sequential inference judgment. Journal of Experimental Psychology, 93, 63-68.

Shanteau, J. (1975). An information-integration analysis of risky decision making. In M. F. Kaplan and S. Schwartz, (Eds.), Human Judgment and Decision Processes. New York: Academic Press, pp 109-137.

Slovic, P. (1975). Choice between equally valued alternatives. Journal of Experimental

Psychology: Human Perception and Performance, 1, 280-287.

Slovic, P., Fischhoff, B., and Lichtenstein, S. (1977). Behavioral decision theory. Annual Review of Psychology, 28, 1-39.

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.

Smolensky, Paul (1986). Neural and conceptual interpretations of parallel distributed vprocessing models. In J.L. McClelland, D.E. Rumelhart, and the PDP Research Group (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume II: Psychological and Biological Models. Cambridge, MA.: MIT Press/Bradford Books.

Tversky, A. (1981). L. J. Cohen, again: On the evaluation of inductive intuitions. The Behavioral and Brain Sciences, 4, 354-356.

Tversky, A., and Kahneman, D. (1971) Belief in the law of small numbers. Psychological Bulletin, 76, 105-110.

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, NJ.: Erlbaum, pp 49-72.

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

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

Wallsten, T.S., Budescu, D.V., Rapoport, A., Zwick, R., and Forsyth, B. (1986). Measuring the vague meanings of probability terms. Journal of Experimental Psychology: General, 115, 348-365.

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.

Wyer, R.S., Jr. (1976) An investigation of the relations among probability estimates. Organizational Behavior and Human Performance, 15, 1-18.

7. Tables.

Table 1. Book Bag and Poker Chip problem

There are two bookbags filled with poker chips. One contains 70% red poker chips and 30% blue chips; the other contains 30% red and 70% blue chips [p(E/H), p(E/~H), p(~E/H), and p(~E/~H].

One will be picked at random [p(H) = .5], and your job will be to guess whether it was the predominantly red or the predominantly blue bag.

A red chip is drawn from the bag [the evidence, E].

What is the probability that it was the predominantly red bookbag [p(H/E)]?

(adapted from Slovic and Lichtenstein, 1971, p 668.)

Table 2. The Cab Problem.

In this city there are only two cab companies, the Blue Cab Company and the Green Cab Company. The Green Cab Company is larger, with 85% of the cabs in the city. [p(H)]

There was a fatal hit and run accident at night. Although the viewing conditions were poor, the only witness identified the cab as blue. [evidence]

The police tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness could correctly identify cabs of each one of the two colors 80% of the time and misidentified them 20% of the time. [p(E/H) and p(E/~H)]

What is the probability that it was a Blue Cab? _____

(Adapted from Tversky and Kahneman, 1982, p 156)

Table 3. The Cab Problem used in this study, with the separate parts identified.

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

In this city there are only two cab companies, the Blue Cab Company and the Green Cab Company. With what you know now, what is the probability that the cab involved in the hit and run accident was from the Blue Cab Company? _____

What is the probability it was from the Green Cab Company? _____

Evidence. There was only one witness to the hit and run accident. The witness identified the cab as blue. With what you know now, what is the probability that it was a Blue Cab? _____

What is the probability that it was from the Green Cab Company? _____

Base rate. The Green Cab Company is larger, with 85% of the cabs in the city. With what you know now, what do you think is the probability that a cab from the Blue Cab Company was the one involved in the accident? _____

What is the probability it was a cab from the Green Cab Company? _____

Reliability. The police were concerned about the accuracy of the witness who saw the accident. They tested the witness's reliability under the same circumstances that existed on the night of the accident and concluded that the witness could correctly identify cabs of each one of the two colors 80% of the time and misidentified them 20% of the time. With what you know now, what is the probability that the cab was a Blue Cab? _____

What is the probability that it was a Green Cab? _____

Table 4. The Doctor Problem.

Introduction. 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 each of two diseases.

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? _____

What is the probability that the patient has hepatitis? _____

Reliability. 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 10%, and is right 90% of the time. That is, when the patient has toxic uremia, the Spock test says so 90% of the time, but it falsely indicates that the patient has hepatitis 10% of the time. Similarly, when the patient has hepatitis, the Spock test will indicate that the disease is toxic uremia about 10% of the time. With what you know now, what is the probability that the patient has toxic uremia? _____

What is the probability the patient has hepatitis? _____

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? _____

What is the probability that the patient has hepatitis? _____

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 75% of people with the symptom of yellow eyes have hepatitis, and only 25% of them have toxic uremia. With what you now know, what is the probability that the patient has toxic uremia? _____

What is the probability the patient has hepatitis? _____

Table 5. The Twins Problem.

Introduction. 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 each 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?_____

What is the probability that it was Paul? _____

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? _____

What is the probability that Paul broke the lamp? _____

Reliability. 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?_____

What is the probability that Paul 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?_____

What is the probability that Paul broke the lamp? _____

Table 6. Mean and median answers at each step of each information presentation order, for each problem.

Cab Problem.

Information Presentation	no info		1st info		2nd info		3rd info		
Order	Mean	Median	Mean	Median	Mean	Median	Mean	Median	N
ber[a,b]	.50	.50	.38	.30	.79	.90	.69	.80	44
bre	.51	.50	.31	.15	.30	.20	.67	.80	40
ebr	.50	.50	.85	.90	.67	.80	.72	.80	42
erb	.50	.50	.85	.90	.76	.80	.61	.68	43
rbe	.50	.50	.49	.50	.31	.15	.67	.80	45
reb	.50	.50	.53	.50	.75	.80	.64	.75	42
All ord	.50	.50	Ir[c]	Ir	Ir	Ir	.67	.80	256

Doctor Problem.

Information Presentation	no info		1st info		2nd info		3rd info		
Order	Mean	Median	Mean	Median	Mean	Median	Mean	Median	N
ber[d]	.49	.50	.27	.25	.93	1.00	.84	.90	41
bre	.48	.50	.30	.25	.29	.25	.82	.90	47
ebr	.50	.50	.95	1.00	.85	.99	.84	.90	43
erb	.48	.50	.96	1.00	.88	.90	.66	.70	43
rbe	.49	.50	.51	.50	.30	.25	.75	.90	43
reb	.49	.50	.53	.50	.86	.90	.68	.75	42
All ord	.49	.50	Ir	Ir	Ir	Ir	.77	.90	259

Twins Problem.

Information Presentation	no info		1st info		2nd info		3rd info		
Order	Mean	Median	Mean	Median	Mean	Median	Mean	Median	N
ber[e]	.49	.50	.26	.20	.52	.50	.44	.50	44
bre	.50	.50	.26	.20	.34	.35	.45	.50	43
ebr	.50	.50	.65	.68	.45	.45	.46	.50	42
erb	.50	.50	.65	.70	.58	.60	.34	.25	43
rbe	.50	.50	.50	.50	.27	.20	.39	.35	42
reb	.49	.50	.49	.50	.57	.60	.37	.40	48
All ord	.50	.50	Ir	Ir	Ir	Ir	.41	.40	262

[a]Information order: b = base rate, e = evidence, r = reliability.
[b]Cab Problem: b = .15, e = 1.0, r = .80.
[c]The overall mean is meaningless in these columns, because the available information at these steps differs across presentation orders.
[d]Doctor Problem: b = .25, e = 1.0, r = .90.
[e]Twins Problem: b = .20, e = 1.0, r = .60.

Table 7. Correct answers for each step, each information presentation order, of the Cab problem.

Information Presentation Order		Amount of information presented: None	One Piece	Two Pieces	Three Pieces
ber	Formula[a]	.5[b]	b	b to e[c]	Bayes'[d]
	Cab	.5	.15	.15-1	.41
	Doctor	.5	.25	.25-1	.75
	Twins	.5	.20	.20-1	.27
bre	Formula	.5	b	b[e]	Bayes'
	Cab	.5	.15	.15	.41
	Doctor	.5	.25	.25	.75
	Twins	.5	.20	.20	.27
ebr	Formula	.5	.5 to e[c]	b to e	Bayes'
	Cab	.5	.5-1	.15-1	.41
	Doctor	.5	.5-1	.25-1	.75
	Twins	.5	.5-1	.20-1	.27
erb	Formula	.5	.5 to e	r[f]	Bayes'
	Cab	.5	.5-1	.80	.41
	Doctor	.5	.5-1	.90	.75
	Twins	.5	.5-1	.60	.27
rbe	Formula	.5	.5[e]	b	Bayes'
	Cab	.5	.5	.15	.41
	Doctor	.5	.5	.25	.75
	Twins	.5	.5	.20	.27
reb	Formula	.5	.5	r	Bayes'
	Cab	.5	.5	.80	.41
	Doctor	.5	.5	.90	.75
	Twins	.5	.5	.60	.27

[a]First row gives general formula for correct answer for all problems; next three rows give specific correct answers for each problem.

[b]Lacking other information, p(H) = .5.

[c]Any answer in this range would be reasonable, depending on reliability (not yet given).

[d]Bayes' Theorem:

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

[e]Reliability has no impact in the absence of evidence.

[f]Application of Bayes' Theorem with a prior of p(H) = .5 yields p(H/E) = p(E/H), the reliability.

Table 8.

Mean absolute deviation of subject's answer from correct answer, at each step of each information presentation order, for each problem.

Cab Problem.

Infor. Pres'n. Order	no info Mean	no info St Dev	1 piece info Mean	1 piece info St Dev	2 pieces info Mean	2 pieces info St Dev	3 pieces info Mean	3 pieces info St Dev
ber	.01	.05	.23	.27	NA[a]	NA	.33	.11
bre	.00	.03	.16	.25	.16	.21	.35	.11
ebr	.00	.00	NA	NA	NA	NA	.32	.11
erb	.00	.00	NA	NA	.05	.09	.27	.14
rbe	.00	.02	.04	.08	.16	.23	.33	.15
reb	.00	.00	.07	.12	.06	.13	.30	.13
Mean	.00	.03	.12	.21	.11	.18	.32	.13

Doctor Problem.

Infor. Pres'n. Order	no info Mean	no info St Dev	1 piece info Mean	1 piece info St Dev	2 pieces info Mean	2 pieces info St Dev	3 pieces info Mean	3 pieces info St Dev
ber	.01	.05	.03	.11	NA	NA	.16	.08
bre	.02	.09	.05	.14	.07	.17	.18	.11
ebr	.00	.00	NA	NA	NA	NA	.15	.08
erb	.02	.08	NA	NA	.02	.07	.20	.17
rbe	.01	.05	.06	.13	.06	.13	.22	.17
reb	.01	.05	.07	.15	.04	.14	.21	.17
Mean	.01	.06	.05	.13	.05	.13	.19	.14

Twins Problem.

Infor. Pres'n. Order	no info Mean	no info St Dev	1 piece info Mean	1 piece info St Dev	2 pieces info Mean	2 pieces info St Dev	3 pieces info Mean	3 pieces info St Dev
ber	.01	.08	.08	.17	NA	NA	.20	.11
bre	.00	.00	.07	.13	.15	.15	.22	.18
ebr	.00	.00	NA	NA	NA	NA	.21	.14
erb	.00	.00	NA	NA	.10	.09	.15	.14
rbe	.00	.00	.01	.04	.07	.17	.15	.13
reb	.01	.07	.03	.09	.08	.13	.17	.14
Mean	.00	.04	.05	.12	.10	.14	.19	.14

[a]"NA" indicates that there is no single correct answer for some combinations of information.

Table 9. Number of subjects with correct answer, after each piece of information, each information presentation order, for each problem.

Cab Problem.

	No information		1 piece info		2 pieces info		3 pieces info	
	Wrong	Right	Wrong	Right	Wrong	Right	Wrong	Right
ber	2	42	26	18	(0)[a]	(44)	44	0
bre	1	39	18	22	30	10	40	0
ebr	0	42	(0)	(42)	(2)	(40)	42	0
erb	0	43	(0)	(43)	11	32	43	0
rbe	1	44	11	34	25	20	45	0
reb	0	42	11	31	10	32	42	0
Total[b]	4	252	66	105	76	94	256	0
			(66)	(190)	(78)	(178)		
% Correct		98.4		61.4		55.3		0.0
				(74.2)		(69.5)		

Doctor Problem.

	No information		1 piece info		2 pieces info		3 pieces info	
	Wrong	Right	Wrong	Right	Wrong	Right	Wrong	Right
ber	2	39	3	38	(0)	(41)	41	0
bre	4	43	7	40	17	30	47	0
ebr	0	43	(0)	(43)	(0)	(43)	41	2
erb	3	40	(0)	(43)	7	36	41	2
rbe	1	42	11	32	12	31	43	0
reb	1	41	10	32	5	37	38	4
Total	11	248	31	142	41	134	251	8
			(31)	(228)	(41)	(218)		
% Correct		95.8		82.1		76.6		3.1
				(88.0)		(84.2)		

Table 9 is continued on next page.

Table 9, continued

Twin Problem.

	No information		1 piece info		2 pieces info		3 pieces info	
	Wrong	Right	Wrong	Right	Wrong	Right	Wrong	Right
ber	1	43	13	31	(1)	(43)	44	0
bre	0	43	14	29	28	15	42	1
ebr	0	42	(1)	(41)	(1)	(42)	42	0
erb	0	43	(1)	(42)	28	15	43	0
rbe	0	42	6	36	11	31	42	0
reb	1	47	7	41	22	26	48	0
Total	2	260	40	137	89	87	261	1
			(42)	(220)	(91)	(171)		
% Correct		99.2		77.4		49.4		0.4
				(84.0)		(65.3)		

[a]For the combinations of information for which there is no single correct answer, the number of subjects answering in the acceptable range is given.

[b]Totals and %s Correct in parentheses include subjects for whom any answer in a given range could be considered correct.

Table 10. The number of subjects who used each of the available answers, given one piece of information.

Cab Problem	Information			TOTAL
	B	E	R	
50/50 or .5	2	10	65	77
Base rate, p(H)	40	0[a]	0[a]	40
Reliability	1[a]	6[a]	8	15
Evidence, 1.0	0[a]	33	0[a]	33
p(~H) from text	12	1[a]	0[a]	13
Original answer[b]	0	0	1	1
1 - reliability	3[a]	0[a]	2	5
Another number	25	35	11	71
TOTAL	83	85	87	255

Doctor Problem	B	E	R	TOTAL
50/50 or .5	1	1	64	66
Base rate, p(H)	76	0[a]	0[a]	76
Reliability	0[a]	14[a]	6	20
Evidence, 1.0	0[a]	44	1[a]	45
p(~H) from text	5	2[a]	0[a]	7
Original answer	3	0	1	4
1 - reliability	1[a]	0[a]	1	2
Another number	2	24	12	38
TOTAL	88	85	85	258

Twins Problem	B	E	R	TOTAL
50/50 or .5	8	31	77	116
Base rate, p(H)	60	1[a]	1[a]	62
Reliability	0[a]	7[a]	6	13
Evidence, 1.0	0[a]	3	0[a]	3
p(~H) from text	4	9[a]	0[a]	13
Original answer	1	0	1	2
1 - reliability	3[a]	0[a]	5	8
Another number	11	34	0	45
TOTAL	87	85	90	262

[a]Answers in these cells could not have been produced by strategy of using an available number, because the number had not been presented yet.

[b]If different from above categories.

Table 11. The number of subjects who used each of the available answers, given two pieces of information.

Cab Problem	Information Presentation Order						TOTAL
	be	br	eb	er	rb	re	
50/50 or .5	5	3	5	1	2	2	18
Base rate, p(H)	1	10	5	0[a]	20	0[a]	36
Reliability	1[a]	0	3[a]	32	2	32	70
Evidence, 1.0	11	0[a]	8	0	0[a]	1	20
p(~H) from text	2	1	7	0[a]	3	0[a]	13
Original answer	0	0	0	0	0	0	0
Previous answer[b]	0	7	7	0	0	1	15
1 - reliability	0[a]	3	0[a]	0	2	0	5
Another number	23	16	5	10	16	6	76
TOTAL	43	40	40	43	45	42	253

Doctor Problem	be	br	eb	er	rb	re	TOTAL
50/50 or .5	1	1	2	1	5	1	11
Base rate, p(H)	1	30	3	0[a]	31	0[a]	65
Reliability	4[a]	2	4[a]	36	0	37	83
Evidence, 1.0	23	0[a]	18	0	0[a]	1	42
p(~H) from text	3	1	2	0[a]	2	0[a]	8
Original answer	0	0	0	0	0	0	0
Previous answer	0	1	6	0	3	0	10
1 - reliability	0[a]	2	0[a]	0	0	1	3
Another number	9	10	8	6	1	1	35
TOTAL	41	47	43	43	42	41	257

Twins Problem	be	br	eb	er	rb	re	TOTAL
50/50 or .5	8	9	7	12	2	14	52
Base rate, p(H)	11	15	12	0[a]	31	0[a]	69
Reliability	3[a]	2	4[a]	15	0	26	50
Evidence, 1.0	1	0[a]	1	0	0[a]	0	2
p(~H) from text	6	1	1	5[a]	3	1[a]	17
Original answer	1	0	0	0	0	0	1
Previous answer	0	2	4	2	0	0	8
1 - reliability	1[a]	7	4[a]	4	1	0	17
Another number	13	7	8	5	5	5	43
TOTAL	44	43	41	43	42	48	261

[a]Answers in these cells could not have been produced by strategy of using an available number, because the number had not been presented yet.

[b]If different from above categories.

Table 12. The number of subjects who used each of the available answers, given three pieces of information.

Cab Problem	Information Presentation Order						TOTAL
	ber	bre	ebr	erb	rbe	reb	
50/50 or .5	2	1	2	3	1	1	10
Base rate, p(H)	3	1	0	5	1	2	12
Reliability	21	22	18	11	13	16	101
Evidence, 1.0	0	1	0	1	1	1	4
Doubt evid, 0	0	0	0	0	0	0	0
p(~H) from text	3	0	3	0	3	3	12
Original answer[a]	0	0	0	0	1	0	1
Previous answer[a]	4	6	1	0	2	1	14
1 - reliability	1	0	0	0	0	0	1
Another number	10	8	18	22	20	18	96
TOTAL	44	39	42	42	42	42	251

Doctor Problem	ber	bre	ebr	erb	rbe	reb	TOTAL
50/50 or .5	1	1	1	2	1	2	8
Base rate, p(H)	0	1	0	4	0	2	7
Reliability	31	36	32	14	25	17	155
Evidence, 1.0	0	1	1	0	2	1	5
Doubt evid, 0	0	0	0	0	0	0	0
p(~H) from text	0	0	2	2	0	4	8
Original answer	0	0	0	0	0	1	1
Previous answer	0	0	0	0	1	1	2
1 - reliability	0	0	0	0	1	1	2
Another number	9	7	7	21	12	13	69
TOTAL	41	46	43	43	42	42	257

Twins Problem	ber	bre	ebr	erb	rbe	reb	TOTAL
50/50 or .5	13	14	12	3	6	8	56
Base rate, p(H)	6	6	4	17	11	8	52
Reliability	5	6	4	2	5	2	24
Evidence, 1.0	0	2	0	0	0	0	2
Doubt evid, 0	0	1	0	0	0	1	2
p(~H) from text	0	0	0	2	0	2	4
Original answer	0	0	0	0	0	0	0
Previous answer	2	5	1	2	2	2	14
1 - reliability	8	2	9	6	5	14	44
Another number	10	7	11	11	13	11	63
TOTAL	44	43	41	43	42	48	261

[a]If different from above categories.

Table 13. Summary of the use of available numbers.

One piece of information.

	Available numbers			Non-available
	Correct	Other	Total	
Cab	58.0%	16.3%	74.3%	25.7%
Doctor	71.7%	10.5%	82.2%	17.8%
Twins	65.3%	12.9%	78.2%	21.8%

Two pieces of information.

	Available numbers			Non-available
	Correct	Other	Total	
Cab	51.0%	17.4%	68.4%	31.6%
Doctor	70.8%	12.5%	83.3%	16.7%
Twins	44.1%	32.6%	76.7%	23.3%

Three pieces of information.

	Available numbers			Non-available
	Reliability	Other	Total	
Cab	40.2%	21.5%	61.7%	38.2%
Doctor	60.3%	13.1%	73.4%	26.8%
Twins	9.2%	66.7%	75.9%	24.1%

Table 14.
Formulas used in categorization scheme.

Addition. Sum of available numbers.

b + pr; b + r; pr + r; c(b) + pr; c(b) + r; b + c(r);
pr + c(r); c(b) + c(r);

Sum of previous answer and most recent information.

la + mri; c(la) + mri; la + c(mri); c(la) + c(mri).

Subtraction. Difference between available numbers.

abs(b-pr); abs(b-r); abs(b-e); abs(pr-r); abs(pr-e); abs(r-e);
abs(c(b)-pr); abs(c(b)-r); abs(c(b)-e); abs(b-c(r)); abs(pr-c(r));
abs(c(r)-e); abs(c(b)-c(r)); abs(b - c(b)); abs (r - c(r));

Difference between previous answer and most recent information.

abs(la - mri); abs(la - c(mri)); abs(c(la) - mri); abs(c(la) - c(mri));
abs(la - c(la)).

Multiplication. Product of available numbers.

b*pr; b*r; pr*r; b*c(r); pr*c(b); pr*c(r); c(b)*c(r); b*c(b);
r*c(b); r*c(r);

Product of previous answer and most recent information.

la*c(la); la*mri; la*c(mri); c(la)*mri; c(la)*c(mri).

Multiplicative triplets.

b*r*pr; b*r*c(r); b*pr*c(r); b*r*c(b); b*c(b)*pr; c(b)*r*c(r);
pr*r*c(b); pr*c(b)*c(r).

Division. Quotient of available numbers.

b/pr; b/r; pr/r; b/c(r); c(b)/c(r); b/c(b); pr/c(b); r/c(b);
pr/c(r); r/c(r);

Quotient of previous answer and most recent information.

la/c(la); la/mri; la/c(mri); c(la)/mri; c(la)/c(mri).

Table 14 is continued on next page.

Table 14, continued.

(A + P)*U. Distributive rule.

(b + pr)*r; (b + r)*pr; (r + pr)*b; (c(b) + pr)*r; (c(b) + r)*pr;
(r + pr)*c(b); (b + pr)*c(r); (b + c(r))*pr; (c(r) + pr)*b;
(c(b) + pr)*c(r); (c(b) + c(r))*pr; (c(r) + pr)*c(b).

A*P + U. Dual-distributive rule.

b*pr + r; b*r + pr; pr*r + b; c(b)*pr + r; c(b)*r + pr;
pr*r + c(b); b*pr + c(r); b*c(r) + pr; pr*c(r) + b; c(b)*pr + c(r);
c(b)*c(r) + pr; pr*c(r) + c(b); pr*r*c(r); b*c(b)*c(r).

Bayes' Theorem. Application of Bayes' Theorem to available numbers.

Using prior of .5: (.5*r)/(.5*r + c(.5)*c(r)) = r;
Using base rate: (b*r)/(b*r + c(b)*c(r));
Using wrong base rate: (c(b)*r)/(c(b)*r + b*c(r));
Using wrong reliability: (b*c(r))/(b*c(r) + c(b)*r);
Using wrong base rate
and wrong reliability: (c(b)*c(r))/(b*r + c(b)*c(r)).

Conventional numbers. Used common probabilities.

.5; .25; .75; .33 or .333; .66 or .67 or .666 or .667; .2; .4; .6;
.8; .1; .3; .5; .7; .9; .05; .95; .02; .98; .01; .99.

---

b = base rate; r = reliability; pr = prior (.5); la = last answer; mri = most recent information; c(x) = complement of x, i.e., (1-x); abs(x) = absolute value of x.

Table 15. Number of answers matching each strategy category, with one piece of information.

Cab problem	Information B	E	R	TOTAL
Number from word problem	12	0	8	20
# implicit in word problem	42	43	65	150
1 - an available #	0	0	2	2
Function of previous answer	1	0	1	2
Sum of available numbers	0	0	1	1
Difference betw available #s	6	0	1	7
Diff, avail # & previous ans	0	0	1	1
Product of available #s	1	0	8	9
Product, avail # & previous ans	3	0	0	3
Quotient of available #s	5	0	0	5
Used conventional probabilities	11	38	0	49
Used numbers in various ranges	3	4	0	7
TOTAL	84	85	87	256

Doctor Problem	B	E	R	TOTAL
Number from word problem	5	0	6	11
# implicit in word problem	77	45	64	186
1 - an available #	0	0	1	1
Function of previous answer	3	1	1	5
Sum of available numbers	0	0	2	2
Difference betw available #s	0	0	3	3
Product of available #s	0	0	3	3
Quotient of avail & prev #s	0	0	1	1
Used conventional probabilities	3	33	2	38
Used numbers in various ranges	0	7	2	9
TOTAL	88	86	85	259

Table 15 is continued on the next page.

Table 15, continued.

Twins Problem	B	E	R	TOTAL
Number from word problem	4	0	6	10
# implicit in word problem	68	34	77	179
1 - an available #	0	0	5	5
Function of previous answer	1	0	1	2
Difference betw available #s	1	0	1	2
Product of available #s	6	0	0	6
Product, avail # & previous ans	6	0	0	6
Used conventional probabilities	1	49	0	50
Used numbers in various ranges	0	2	0	2
TOTAL	87	85	90	262

Table 16. Number of answers matching strategy category, with two pieces of information.

Cab Problem	Information Presentation Order						
	be	br	eb	er	rb	re	TOTAL
Number from word problem	2	1	7	32	5	32	79
# implicit in word problem	17	13	18	1	22	3	74
1 - an available #	0	3	0	0	2	0	5
Function of previous answer	1	7	10	0	0	1	19
Sum of available numbers	2	0	0	3	3	1	9
Sum of avail # & previous ans	0	0	1	0	0	0	1
Difference betw available #s	0	3	0	3	2	1	9
Diff, avail # & previous ans	0	0	1	1	3	0	5
Product of available #s	0	7	1	1	6	2	17
Product, avail # & previous ans	0	1	0	0	1	0	2
Quotient of available #s	1	4	1	0	1	0	7
Used conventional probabilities	18	0	2	0	0	1	21
Used numbers in various ranges	3	1	1	2	0	1	8
TOTAL	44	40	42	43	45	42	256

Doctor Problem	be	br	eb	er	rb	re	TOTAL
Number from word problem	3	3	2	36	2	37	83
# implicit in word problem	25	31	23	1	36	2	118
1 - an available #	0	2	0	0	0	1	3
Function of previous answer	0	1	9	0	4	1	15
Sum of available numbers	0	1	0	0	0	1	2
Difference betw available #s	0	4	0	3	0	0	7
Diff, avail # & previous ans	0	1	2	1	0	0	4
Product of available #s	0	2	0	0	0	0	2
Quotient of available #s	0	1	0	0	0	0	1
Used conventional probabilities	9	0	4	0	0	0	13
Used numbers in various ranges	4	1	3	2	1	0	11
TOTAL	41	47	43	43	43	42	259

Table 16 is continued on the next page.

Table 16, continued.

Twins Problem	be	br	eb	er	rb	re	TOTAL
Number from word problem	6	3	1	15	3	26	54
# implicit in word problem	20	24	20	12	33	16	125
1 - an available #	0	7	0	4	1	0	12
Function of previous answer	1	2	7	5	0	0	15
Sum of available numbers	2	0	1	0	0	2	5
Difference betw available #s	6	1	6	0	3	0	16
Diff, avail # & previous ans	0	1	0	0	1	0	2
Product of available #s	1	2	2	2	0	0	7
Quotient of available #s	1	0	1	2	0	1	5
Used conventional probabilities	5	1	2	1	0	3	12
Used numbers in various ranges	2	2	2	2	1	0	9
TOTAL	44	43	42	43	42	48	262

Table 17. Number of answers matching each strategy category, with three pieces of information.

Cab Problem	Information Presentation Order						
	ber	bre	ebr	erb	rbe	reb	TOTAL
Number from word problem	24	22	21	11	16	19	113
# implicit in word problem	5	3	2	9	3	4	26
1 - an available #	1	0	0	0	0	0	1
Function of previous answer	4	7	1	1	6	1	20
Sum of available numbers	0	2	4	6	4	5	21
Sum of avail # & previous ans	0	0	0	1	0	0	1
Difference betw available #s	1	0	6	4	0	3	14
Diff, avail # & previous ans	1	0	0	0	0	3	4
Product of available #s	3	0	1	3	6	2	15
Product, avail # & previous ans	0	0	0	1	0	0	1
Quotient of available #s	3	0	2	3	1	2	11
Combination: (A+P)*U	0	1	1	1	1	0	4
Used conventional probabilities	0	4	1	1	4	0	10
Used numbers in various ranges	2	1	3	2	4	3	15
TOTAL	44	40	42	43	45	42	256

Doctor Problem	ber	bre	ebr	erb	rbe	reb	TOTAL
Number from word problem	31	36	34	16	25	21	163
# implicit in word problem	1	3	2	6	3	5	20
1 - an available #	0	0	0	0	1	1	2
Function of previous answer	0	1	0	0	2	2	5
Sum of available numbers	3	2	1	3	0	1	10
Difference betw available #s	3	2	0	5	4	3	17
Diff, avail # & previous ans	2	0	0	1	0	0	3
Product of available #s	1	1	0	2	3	2	9
Product, avail # & previous ans	0	0	1	0	0	0	1
Quotient of available #s	0	0	0	1	0	0	1
Used conventional probabilities	0	2	3	4	1	5	15
Used numbers in various ranges	0	0	2	5	4	2	13
TOTAL	41	47	43	43	43	42	259

Table 17 is continued on the next page.

Table 17, continued.

Twins Problem	ber	bre	ebr	erb	rbe	reb	TOTAL
Number from word problem	5	6	4	4	5	4	28
# implicit in word problem	19	23	16	20	17	17	112
1 - an available #	8	2	9	6	5	14	44
Function of previous answer	2	5	2	2	2	2	15
Sum of available numbers	3	2	1	2	0	2	10
Difference betw available #s	1	1	2	3	8	4	19
Diff, avail # & previous ans	1	0	0	4	0	1	6
Product of available #s	3	2	1	0	0	0	6
Product, avail # & previous ans	1	0	1	0	0	0	2
Quotient of available #s	0	0	3	0	3	2	8
Used conventional probabilities	0	0	0	1	0	0	1
Used numbers in various ranges	1	2	3	1	2	2	11
TOTAL	44	43	42	43	42	48	262

Table 18. Summary. Number of answers matching each class of strategy, at each step.

	PROBLEM								
	Cab			Doctor			Twins		
Amount of info	1	2	3	1	2	3	1	2	3
Class of strategy used									
Available number	172	158	140	198	204	185	194	191	184
Previous answer	2	19	20	5	15	5	2	15	15
Arithmetic oper.	26	50	71	9	16	4	14	35	51
Conventional prob.	49	21	10	38	13	15	50	12	1
Other	7	8	15	9	11	13	2	9	11
Total	256	256	256	259	259	259	262	262	262

Table 19. Range of ambiguity in categorization procedure for the arithmetic operations and the use of conventional probabilities, in contrast with the use of available numbers. Cab problem data, three pieces of information.

OPERATION	ANALYSIS[a]	Category to which answer is assigned by each analysis procedure. N = 256. Available Number	Arithmetic Operation	Other
Addition	Av	160	22	74
	Op	-	43	213
Subtraction	Av	160	18	78
	Op	-	170	86
Multiplication	Av	160	16	80
	Op	-	42	214
Division	Av	160	11	85
	Op	-	43	213
(A+P)*U, A*P+U, or Bayes' Theorem	Av	160	4	92
	Op	-	26	230
Conventional probabilities	Av	160	10	86
	Op	-	189	67

[a]Analysis Av assigns ambiguous answers to the strategy of using an available number (or to an arithmetic operation strategy earlier on the list); Analysis Op assigns them to the arithmetic operation strategy.

Table 20. Transcript of pilot subject working on a version of the Cab Problem.

Subject is given base rate of 25% blue cabs, and is asked, "With what you know now, what do you think is the probability that a cab from the Blue Cab Company was the one involved in the accident?"

"I'd say about a 25% chance... because... there'd be 25 cabs for every 75 of the green ones."

Subject is told that the witness identified the cab as blue.

Um.... I'd probably make it higher, about 40, maybe because even though the percent of cabs is only 25%, ... if someone thought... it was blue, then they obviously probably asked him if he thought it was green or blue and he said blue and it raised the percentage, but I still don't think it would raise it that high, because it was at night. Green and blue is hard to tell [apart], because... the viewing conditions were poor."

Subject is told that the witness's reliability is 70% correct.

"Um.... I'd raise the probability again since 70 is a pretty high number, so I'd think that it'd be, I'd probably make it a number between 40 and 70. Which would be about... 60... between 60... I'd say about 60% chance that [it was a blue cab]."

Table 21. Number of subjects whose answers fall within range of available numbers, at each step of each problem.

Amount of Information. Problem.	Number of Subjects: Below Range N	In Range N	In Range %	Above Range N
None				
Cab	1	252	98.4%	3
Doctor	10	248	95.8%	1
Twin	2	260	99.2%	0
One piece				
Cab	13	226	88.3%	17[a]
Doctor	10	240	92.7%	9
Twin	12	246	93.9%	4
Two pieces				
Cab	22	231	90.2%	3[a]
Doctor	18	241	93.1%	0
Twin	18	241	92.0%	3
Three pieces				
Cab	8	248	96.9%	-[a]
Doctor	17	241	93.1%	-
Twin	10	252	96.2%	-

[a]When the evidence information (e = 1.0) has been presented, it is not possible for subjects to answer above the range. This is true for one third of subjects after 1 piece of information, for two thirds after two pieces, and for all subjects after 3 pieces of information.

Table 22. Estimates for relative weights of Base Rate, Evidence, and Reliability, based on the "Nearness" Assumption.

		Comparisons					
		Base Rate and Reliability		Evidence and Reliability		Base Rate and Evidence	
Problem	Analysis	B	R	R	E	E	B
Cab	I=1, no .5	.62	.38	.56	.44	.64	.36
	I=1, .5	.39	.24	.39	.31	.38	.22
	I=range, no .5	.75	.25	.64	.36	.68	.32
	I=range, .5	.44	.14	.48	.27	.40	.18
	Pattern	B > R		R > E		E > B	
Doctor	I=1, no .5	.71	.29	.53	.47	.71	.29
	I=1, .5	.45	.18	.39	.35	.48	.19
	I=range, no .5	.94	.06	.56	.44	.85	.15
	I=range, .5	.56	.04	.49	.38	.58	.10
	Pattern	B > R		R > E		E > B	
Twins	I=1, no .5	.56	.44	.63	.37	.41	.59
	I=1, .5	.37	.30	.39	.23	.22	.32
	I=range, no .5	.72	.28	.87	.13	.36	.64
	I=range, .5	.48	.18	.49	.07	.18	.32
	Pattern	B > R		R > E		E < B	

Table 23. Within-subjects comparisons of the impact of base rate information.

Comparison 1st	2nd	M(1st)	M(2nd)	T	p	N
Cab Problem						
0	b	.50	.34	5.76	.000	84
e	eb	.85	.67	4.12	.000	42
r	rb	.49	.31	4.57	.000	45
er	erb	.76	.61	4.28	.000	43
re	reb	.75	.64	3.61	.001	42
(er)	(er)b	.75	.63	5.60	.000	85
Doctor Problem						
0	b	.48	.29	12.82	.000	88
e	eb	.95	.85	3.52	.001	43
r	rb	.51	.30	10.32	.000	43
er	erb	.88	.66	5.76	.000	43
re	reb	.86	.68	3.86	.000	42
(er)	(er)b	.87	.67	6.66	.000	85
Twins Problem.						
0	b	.49	.26	14.05	.000	87
e	eb	.65	.45	6.65	.000	42
r	rb	.50	.27	8.69	.000	42
er	erb	.58	.34	8.69	.000	43
re	reb	.57	.37	7.52	.000	48
(er)	(er)b	.57	.36	11.42	.000	91

Table 24. Between-subject comparisons of the impact of base rate information.

Cab Problem.

Comparison 1st	2nd	M(1st)	M(2nd)	T	p	N1	N2
er	ber	.76	.69	1.94	.055	43	44
er	ebr	.76	.72	1.50	.137	43	42
re	bre	.75	.67	1.67	.100	42	40
re	rbe	.75	.67	1.68	.097	42	45
e	eb	.85	.67	3.24	.002	43	42
e	be	.85	.79	1.56	.122	85	44
r	rb	.53	.31	5.38	.000	42	45
r	br	.51	.30	7.09	.000	87	40
0	b	.50	.34	8.01	.000	172	84

Doctor Problem

Comparison 1st	2nd	M(1st)	M(2nd)	T	p	N1	N2
er	ber	.88	.84	1.53	.130	43	41
er	ebr	.88	.84	1.54	.128	44	43
re	bre	.86	.82	1.14	.258	42	47
re	rbe	.86	.75	2.40	.019	42	43
e	eb	.96	.85	2.87	.005	43	43
e	be	.96	.93	1.42	.157	86	41
r	rb	.53	.30	7.08	.000	42	43
r	br	.52	.29	7.71	.000	85	47
0	b	.49	.29	18.48	.000	171	88

Twins Problem.

Comparison 1st	2nd	M(1st)	M(2nd)	T	p	N1	N2
er	ber	.58	.44	4.51	.000	43	44
er	ebr	.58	.46	3.49	.001	43	42
re	bre	.57	.45	3.03	.003	48	43
re	rbe	.57	.39	5.42	.000	48	42
e	eb	.65	.45	4.45	.000	43	42
e	be	.65	.52	3.42	.001	85	44
r	rb	.49	.27	7.66	.000	48	42
r	br	.49	.34	7.64	.000	90	43
0	b	.50	.26	19.02	.000	175	87

Table 25. Within-subject comparisons of the impact of reliability information.

Comparison 1st	2nd	M(1st)	M(2nd)	T	p	N
Cab Problem.						
eb	ebr	.67	.72	-1.13	.266	42
be	ber	.79	.69	2.84	.007	44
e	er	.85	.76	3.42	.001	43
b	br	.31	.30	0.37	.714	40
0	r	.50	.51	-0.71	.482	87
Doctor Problem.						
eb	ebr	.85	.84	.39	.697	43
be	ber	.93	.84	2.82	.007	41
e	er	.96	.88	6.16	.000	43
b	br	.30	.29	.38	.709	47
0	r	.49	.52	-1.73	.088	85
Twins Problem.						
eb	ebr	.45	.46	-0.23	.821	42
be	ber	.52	.44	2.31	.026	44
e	er	.65	.58	2.49	.017	43
b	br	.26	.34	-4.09	.000	43
0	r	.49	.49	.45	.657	90

Table 26. Between subjects comparisons of the impact of reliability information.

Cab Problem.

Comparison 1st	2nd	M(1st)	M(2nd)	T	p	N1	N2
eb	reb	.67	.64	.40	.687	42	42
eb	erb	.67	.61	.94	.351	42	43
be	rbe	.79	.67	2.42	.018	44	45
be	bre	.79	.67	2.36	.021	44	40
e	re	.85	.75	3.45	.001	85	42
e	er	.85	.76	3.21	.002	42	43
b	br	.38	.30	1.40	.165	44	40
b	rb	.34	.31	.70	.486	84	45
0	r	.50	.51	-0.91	.364	169	87

Doctor problem.

Comparison 1st	2nd	M(1st)	M(2nd)	T	p	N1	N2
eb	reb	.85	.68	3.17	.002	43	42
eb	erb	.85	.66	3.72	.000	43	43
be	rbe	.93	.75	3.61	.001	41	43
be	bre	.93	.82	2.72	.008	41	47
e	re	.96	.86	4.47	.000	86	42
e	er	.95	.88	3.92	.000	43	43
b	br	.27	.29	-0.66	.511	41	47
b	rb	.29	.30	-0.72	.472	88	43
0	r	.49	.52	-2.33	.021	174	85

Twins problem.

Comparison 1st	2nd	M(1st)	M(2nd)	T	p	N1	N2
eb	reb	.45	.37	1.69	.094	42	48
eb	erb	.45	.34	2.26	.027	42	43
be	rbe	.52	.39	2.60	.011	44	42
be	bre	.52	.45	1.24	.219	44	43
e	re	.65	.57	2.82	.005	85	48
e	er	.65	.58	2.42	.018	42	43
b	br	.26	.34	-2.06	.042	44	43
b	rb	.26	.27	-0.38	.702	87	42
0	r	.50	.49	.73	.469	172	90

Table 27.
Production systems representing the normative response strategy and the most common response strategy.

	Condition	Action		
		Normative	Typical subjects	
Rule			Cab and Doctor Problems	Twins Problem
A	Null[a]	estimate b	estimate b	estimate b
B	b	b	b	b
C	e	estimate r estimate b	e	estimate b
D	b and e	estimate r	e	b
E	e and r	estimate b	r	r
F	e, r, and b	Apply Bayes' Theorem	~[b]	.5
G	Query: estimate r no r info	Make best guess adopt r = best guess	~	~
H	Query: estimate b no b info	adopt b = .5	adopt b = .5	adopt b = .5

[a] In addition to the conditions listed, each production has the condition "Query for p(H)".

[b] ~ indicates that there is no rule in the Typical Subject production system corresponding to the rule in the normative production system.

Table 28.
Sequence of production application for normative and typical subject production systems.

	Normative		Typical Subjects			
			Cab and Doctor Problems		Twins Problem	
Given Info.	Right Answer	Rule Sequence	Most common Answer	Rule Sequence	Most common Answer	Rule Sequence
0	.5	AHB	.5	AHB	.5	AHB
e	.5 to e	CGEHF or CHDGF	e	C′	.5	C′′
r	.5	AHB	.5	AHB	.5	AHB
b	b	B	b	B	b	B
(er)	r	EHF	r	E′	r	E′
(eb)	b to e	DGF	e	D′	b	D′′
(rb)	b	B	b	B	b	B
(erb)	BT	F	r	E′	.5	E′′

Table 29. Implicit reliabilities of subjects in various conditions in which reliability information was not given.

Problem and condition			Answer p(H/E)	Base rate	Implicit reliability
Cab	e	Mean	.85	.50[a]	.85
		Median	.90	.50	.90
		Mode	1.00	.50	1.00
	be	Mean	.79	.15	.96
		Median	.90	.15	.98
		Mode	1.00	.15	1.00
	eb	Mean	.67	.15	.92
		Median	.80	.15	.96
		Mode	1.00	.15	1.00
Doctor	e	Mean	.96	.50[a]	.96
		Median	1.00	.50	1.00
		Mode	1.00	.50	1.00
	be	Mean	.93	.25	.98
		Median	1.00	.25	1.00
		Mode	1.00	.25	1.00
	eb	Mean	.85	.25	.94
		Median	.99	.25	.997
		Mode	1.00	.25	1.00
Twins	e	Mean	.65	.50[a]	.65
		Median	.69	.50	.69
		Mode	.50	.50	.50
	be	Mean	.52	.20	.81
		Median	.50	.20	.80
		Mode	.20	.20	.50
	eb	Mean	.45	.20	.77
		Median	.45	.20	.77
		Mode	.20	.20	.50

[a]When base rate information is not given, b = .5 is assumed.

Table 30. Mean responses on Twins problem, Incoherent and Coherent Versions, for each Information Presentation Order Condition.

Incoherent version. N = 221.

Condition	None	One	Two	Three	N
ber	.49	.26	.50	.43	38
bre	.50	.27	.36	.48	37
ebr	.50	.65	.47	.47	36
erb	.50	.65	.57	.33	37
rbe	.50	.50	.26	.40	36
reb	.50	.50	.59	.39	37

Coherent version. N = 41.

	None	One	Two	Three	N
ber	.50	.31	.61	.46	6
bre	.50	.19	.18	.26	6
ebr	.50	.68	.32	.39	6
erb	.50	.63	.61	.43	6
rbe	.50	.50	.35	.28	6
reb	.45	.43	.51	.31	11

Table 31. T-Tests of differences between means of each condition, Coherent and Incoherent Versions of Twin Problem.

Condition	Version: Incoherent Mean	Incoherent N	Coherent Mean	Coherent N	t	p
0	.50	221	.49	41	1.34	.181
b	.26	75	.25	12	.30	.766
e	.65	73	.65	12	-.12	.904
r	.50	73	.45	17	2.63	.010*
(be)	.49	74	.46	12	.33	.744
(br)	.31	73	.26	12	.96	.342
(er)	.58	74	.54	17	.91	.364
be	.50	38	.61	6	-.84	.406
ber	.43	38	.46	6	-.36	.722
br	.36	37	.18	6	2.83	.007*
bre	.48	37	.26	6	2.36	.023*
eb	.47	36	.32	6	1.53	.135
ebr	.47	36	.39	6	.96	.340
er	.57	37	.61	6	-.65	.522
erb	.33	37	.43	6	-1.08	.286
rb	.26	36	.35	6	-1.22	.228
rbe	.40	36	.28	6	1.62	.113
re	.59	37	.51	11	1.54	.132
reb	.39	37	.31	11	1.31	.197

Table 32. Comparison between problems of the effect of evidence in the absence of reliability information.

Evidence alone.

Problem	Condition					
	No inf	e	T	p	N	Effect
Doctor	.49	.96	-41.84	.000	86	.47
Cabs	.50	.85	-18.65	.000	85	.35
Twins Nonsens	.50	.65	-7.68	.000	73	.15
Twins No Nons	.50	.65	-4.96	.000	12	.15

Evidence following base rate.

Problem	Condition					
	b	be	T	p	N	Effect
Doctor	.27	.93	-22.66	.000	41	.66
Cabs	.38	.79	-8.70	.000	44	.41
Twins Nonsens	.26	.50	-5.81	.000	38	.24
Twins No Nons	.31	.61	-2.24	.075	6	.30

Table 33. Strategy Stability Analysis. Comparison between Cab and Doctor Problems.

Complex categorization scheme.

Category	Expected number	Observed number	
base rt, ign evidence	.3	1	
basert * reliability	.1	2	
comp(br) * reliability	.1	2	
within +/- .05 of BT	.0	0	
compl(last answer)	.0	0	
between pr & BT	.3	0	
compl(reliability)	.0	0	
evidence, ign baserate	.1	0	
last answer	.1	0	
ans < prior & basert	.1	0	
stuck with .5	.3	2	
reliability	59.1	71	
reliab < ans < evidenc	.2	1	
prior (other than .5)	.0	0	
all the above cats	60.7	79	chi-squared = 5.71
other categories	204.3	186	p = .025

Simple categorization scheme.

Category	Expected number	Observed number	
< baserate	.9	3	
baserate	.3	1	
br to .5	.6	1	
.5	.3	2	
.5 to rel	11.2	17	
reliability	59.4	71	
rel to 1	.4	1	
certainty, 1	.1	0	
above cats	73.2	96	Chi-squared = 8.39
other cats	177.8	155	p < .005

Table 34. Strategy Stability Analysis. Comparison between Cab and Twins Problems.

Complex categorization scheme.

Category	Expected number	Observed number	
base rt, ign evidence	2.4	2	
basert * reliability	.1	0	
comp(br) * reliability	.0	0	
within +/- .05 of BT	.0	0	
compl(last answer)	.0	0	
compl(reliability)	.2	0	
evidence, ign baserate	.0	0	
between prior and BT	.1	0	
last answer	.3	0	
ans < prior & basert	.1	1	
stuck with .5	2.1	4	
prior < ans < reliab	.7	0	
prior * reliability	.6	0	
reliability	9.1	14	
reliab < ans < evid	1.0	2	
prior (other than .5)	.0	0	
all the above cats	16.7	23	Chi-squared is NS
other categories	248.3	242	

Simple categorization scheme.

Category	Expected number	Observed number	
< baserate	.4	2	
baserate	2.4	2	
br to BT	.7	5	
BT to .5	1.5	3	
.5	2.1	4	
.5 to rel	1.0	1	
reliability	9.5	14	
rel to 1	2.5	3	
certainty, 1	.3	0	
above cats	20.4	34	Chi-squared = 5.83
other cats	233.6	220	p < .025

Table 35. Strategy Stability Analysis. Comparison between Doctor and Twins Problems.

Complex categorization scheme.

Category	Expected number	Observed number	
base rt, ign evidence	1.4	1	
basert * reliability	.0	0	
comp(br) * reliability	.0	1	
within +/- .05 of BT	.0	1	
compl(last answer)	.0	0	
compl(reliability)	.4	0	
between prior & BT	.5	1	
evidence, ign baserate	.0	0	
last answer	.1	0	
ans < prior & basert	.3	0	
stuck with .5	1.7	2	
reliability	14.0	14	
reliab < ans < evidence	.2	0	
prior (other than .5)	.0	0	
all the above cats	18.6	20	Chi-squared is NS
other categories	246.4	245	

Simple categorization scheme.

Category	Expected number	Observed number	
< baserate	.6	0	
baserate	1.4	1	
br to .5	2.3	3	
.5	1.7	2	
.5 to rel	.7	1	
reliability	14.4	14	
rel to 1	.2	0	
certainty, 1	.0	0	
above cats	21.3	21	
other cats	235.7	236	Chi-squared is NS.

Table 36. Correlations between Accuracy (absolute deviation) and various subject variables.

		Exp w/ Problem	Year in college	Time on questnr	Sems col math	Sems col stats
Cab	0	-.10	-.02	-.03	.00	-.02
	1	.04	-.01	.01	-.11	-.03
	2	.02	.02	-.09	-.09	-.07
	3	.07	.10	-.08	.05	.00
Doc	0	.10	.05	.06	-.03	-.04
	1	-.04	-.15*	-.15*	-.10	-.08
	2	-.04	-.01	-.05	-.06	-.07
	3	.03	.00	.19**	.01	.00
Twin	0	-.08	.04	-.02	.06	-.02
	1	-.05	.07	.00	.03	.12
	2	.10	-.04	-.13	-.12	-.08
	3	.09	.01	-.10	-.08	-.08

Table 37. Mean probability that subjects' guesses would be correct if forced to choose.

	Amount of information						p(correct) if chose in accord with Bayes' Theorem answer			
	1	N	2	N	3	N	prob	Dif #3	P(diab)	Dif #3
Cab	.593	174	.742	170	.447	256	.59	.143	.41	.037
Doctor	.607	175	.788	175	.681	259	.75	.069	.25	.431
Twin	.619	179	.634	176	.591	262	.73	.139	.27	.321

Table 38.
Number and proportion of subjects having each possible probability of being correct for each problem, each amount of information.

		Problem								
		Cab			Doctor			Twins		
		Value	N	Prop	Value	N	Prop	Value	N	Prop
Amt	One	.15	18	.103	.25	6	.034	.20	4	.022
of		.50	92	.529	.50	88	.503	.50	100	.559
Info		.85	64	.368	.75	81	.463	.80	75	.419
	Two	.15	13	.076	.10	1	.004	.20	6	.034
		.20	4	.024	.25	6	.034	.40	8	.045
		.50	8	.047	.50	8	.046	.50	37	.210
		.80	78	.459	.75	78	.446	.60	57	.324
		.85	67	.394	.90	82	.469	.80	68	.386
	Three	.41	198	.773	.25	32	.124	.27	51	.195
		.50	10	.039	.50	8	.031	.50	56	.214
		.59	48	.188	.75	219	.846	.73	155	.592