# Structuring as an Aid to Performance in Base-Rate Problems

Sarah Lichtenstein and Donald MacGregor

Decision Research

A Branch of Perceptronics, Inc.

1201 Oak Street

Eugene, OR

97401

Decision Research Report 84-16

Abstract

Four groups of college students were each given two base-rate problems. Three of the groups were given an aid with the first problem: (a) An instruction to list factors or aspects that were relevant to solving the problem, (b) a fill-in-the-blank algorithm that provided the correct solution, or (c) a seven-page tutorial that explained base-rate problems and showed how to solve them using a 2 x 2 table. No aid was provided for the second problem. The control group replicated previous findings in disregarding the base-rate information. The "list factors" group showed no improvement over the control group. The algorithm group showed distinctly better performance for the first problem but were the same as the control group for the second problem. The tutorial group did best: 42% of answers to the first problem and 31% of answers to the second problem were within + .10 of the correct answer. An error analysis identified a conceptual weakness in the tutorial; a high rate of arithmetic errors was also found. College students appear to lack the knowledge needed to solve base-rate problems but they can be taught this knowledge relatively easily.

## Structuring as an Aid to Performance in Base-Rate Problems

A frequently-studied class of inference problems requires the combination of two kinds of probabilistic information, base-rate information, that is, information about the population of events, and diagnostic information, that is, information about the specific event being considered. The base-rate fallacy is the tendency for people to disregard base rates when given these inference problems. For example, consider the following story problem:

> Two companies operate in a given city, the Blue and the Green (according to the color of cab they run). Eighty-five percent of the cabs in the city are Blue and the remaining 15% are Green. A cab was involved in a hit-and-run accident at night. A witness later identified the cab as a Green cab. The court tested the witness' ability to distinguish between Blue and Green cabs under nighttime visibility conditions. It found that the witness was able to identify each color correctly about 80% of the time, but confused it with the other color about 20% of the time.
>
> What do you think are the chances that the errant cab was indeed Green, as the witness claimed?
>
> (Bar-Hillel, 1980, pp. 211-212).

In response to this problem, which is becoming something of a classic, most subjects answer 80%. Similar responses have been found for story problems that are structurally similar but have

different cover stories (e.g., Lyon & Slovic, 1976). The normatively correct answer, derivable by Bayes' Theorem, is a probabilistic merging of both pieces of information provided in the story, resulting in a probability of .41. The subjects' response of .80 indicates a reliance on the diagnostic information given in the story (here, the witness' testimony) and a disregard for the base-rate information (here, the relative number of each color of cab in the city).

Subjects do not always disregard base rates. Research has suggested that they do so only when they believe that the base-rate information is not relevant (Bar-Hillel, 1980). Such information can be made to seem more relevant, for example, by changing, in the above story, the information "85% of the cabs in the city are Blue" to "85% of the cab accidents in the city involve Blue cabs" (Tversky & Kahneman, 1980). This wording apparently evoked a causal link between the population of cabs and the accident being considered. This causal connection heightened the apparent relevance of the base rate (see also Ajzen, 1977).

Most of the research on the base-rate fallacy has focused on variations in the stories, rather than on changing the subjects (Bar-Hillel, 1983; Tversky & Kahneman, 1982). In contrast, the focus of the present study was to explore the effect of different kinds of aids that might help the subjects overcome the base-rate fallacy. One approach that has been tried is to present the subjects with experience, via slides sequentially presenting the

population of cases to the subjects (Christensen-Szalanski & Beach, 1982). That approach was found to be effective, but has been criticized on the grounds that the subjects did not have to integrate the two pieces of information in such a procedure; they could simply count the relative frequency of the desired co-occurence (Beyth-Marom & Arkes, 1983). Moreover, generalization of the improvement to other problems was not tested.

A class of aids called "focusing techniques" has been explored by Fischhoff and his colleagues (Fischhoff, Slovic & Lichtenstein, 1979; Fischhoff & Bar-Hillel, 1984). This approach uses instructions (e.g., "If you only knew the proportion of Green cabs in the city, what would you think is the probability that the cab was Green?") or problem variations (e.g., presenting the same subject with three cab problems, in which the proportion of Green cabs was first 2%, then 98%, then 15%) to focus the subject's attention on the base-rate information. These aids did improve performance, in the sense that the median response was closer to the optimal answer. Unfortunately, they were equally effective in changing subjects' responses to two other problems which were superficially like the cab problem but for which it is optimal to disregard the "base-rate" information. For these problems, performance was worse using the aid. This result suggests that the focusing techniques used in that research did not improve the quality of subjects' thinking about the problems; rather, they created demand characteristics that led the subjects to different responses.

The present paper explores the effectiveness of three aids. The first is a simple instruction to list factors that might be relevant in answering the question. The second is an algorithm; subjects were given the correct process to follow to solve the problem, in the form of a fill-in-the-blanks algorithm; they were not told, however, why this set of calculations was correct. Both these aids have been shown to be effective in a task of estimating unstructured uncertain quantities such as "How many cigarettes were sold in the U.S. last year?" (MacGregor, Lichtenstein & Slovic, 1984). In that study, the performance of subjects given the algorithm was greatly superior to that of a control group; even the "List Factors" group showed some improvement. Presumably, the algorithm, and, to a lesser extent, the "List Factors" instruction, helped the subjects to access and organize their knowledge.

The algorithms previously used required the subjects to estimate some quantities; for example, in the Cigarette algorithm subjects had to estimate the population of the U.S., the proportion who smoke, and the average number of cigarettes a smoker smokes in one day. In contrast, the algorithm for a base-rate problem requires no estimation. To use it, one need only extract from the problem the necessary information, put it in the appropriate spaces, and correctly follow the instructions for arithmetic manipulations on the numbers. Thus, we would expect radical improvement in performance when the algorithm is available.

One difficulty with the correct solution of base-rate problems is that it requires understanding of a moderately complex integration rule. Even an approximation to the correct answer requires an understanding that the two pieces of information need to be played off against each other, one indicating that the desired probability is high, the other that it is low. The algorithm here used, although it does lead to the correct answer, may not illuminate any understanding of the integration process involved. Without such understanding, subsequent performance would be expected to return to unaided levels. To test this conjecture, we presented each subject a second, similar base-rate problem without an algorithm.

For our final aid, we wrote a lengthy tutorial in which we tried to explain both how to do base-rate problems and why our approach was correct. Our goal was to teach the solution to base-rate problems so that subjects would understand the process involved.

## Method

Subjects. The subjects were 305 paid volunteers who responded to ads in the University of Oregon student newspaper. The present tasks were completed along with several other unrelated paper-and-pencil tasks in a one- to two-hour period. Except as noted below, all subjects were run in groups of 30 to 60 people in a large university classroom.

Design. All subjects were given two base-rate problems, here called the Lightbulb problem (adapted from Lyon & Slovic, 1976) and the Dyslexia problem; both are shown in Table 1. Approximately half the subjects received the Lightbulb problem first; the others received the Dyslexia problem first. The two administrations were separated by two unrelated tasks. For all subjects, the second problem was presented in its Control form, the form shown in Table 1. The first problem was presented in four different forms:

1. Control. The Control form was given to 41 subjects.

2. List. The List form was given to 86 subjects. In the List form, after the problem was presented, the instructions read:

> Before answering the question, we would like you to list the things one should consider in answering this question. These things could be a list of factors or components that would be useful in arriving at an answer or they could be ways for going about arriving at an answer. Make your list here:
>
> [seven blank lines]
>
> Now, answer the question:
>
> "What is the probability that this bulb is really defective? [the child really has dyslexia]?
>
> You can probably give a good estimate if you think hard and carefully.
>
> Answer _____

---

Insert Table 1 about here

---

3. Algorithm. The Algorithm form, given to 76 subjects, started with this instruction:

> In this task we would like you to work through a problem by carefully following a number of detailed steps. First, you will read through the problem. Then, you will follow a series of steps, some that ask you to pull information directly from the problem itself, and others that ask you to carry out basic arithmetic. Please follow all the directions carefully. Pay special attention to the accuracy of your arithmetic. This is not a test of your ability to do arithmetic, but accuracy of computation is essential to what we are asking you to do.
>
> [The problem followed.]

After the problem was an algorithm composed of thirteen steps, as shown for the Lightbulb problem in Table 2. On the page following the algorithm, two additional questions were asked:

> Do you think the answer in (M) is a sensible answer to the question, "What is the probability that this lightbulb is really defective [the child really has dyslexia]? Yes ___ No ___
>
> If you answered No, what do you think is a sensible answer? _____

---

Insert Table 2 about here

---

4. Tutorial. The Tutorial form, given to 102 subjects, was a six page, single spaced essay. The seventh page presented the problem with space to work it and a summary of the seven steps to solution discussed in the essay. The last page asked, "Does your answer seem sensible to you? Yes___ No___." However, unlike the Algorithm instructions, a more sensible answer was not requested. Instead, subjects responding "No" were urged to:

> . . . review the steps above. You may have made an error in following the procedure or in doing the arithmetic. Check for errors and correct any you find. OR it may be that your intuitions are wrong and the procedure is correct. Think again about the importance of taking into account both the population information and the specific information.

The tutorial, shown in the Appendix, was based on an approach using 2 x 2 tables rather than Bayes' Theorem, in accordance with Shaughnessy's (1983) view that 2 x 2 tables "help people focus on the restricted sample space which plays so vital a role in conditional probability problems" (p. 344; emphasis in original). It was an expansion of the explanation of base-rate problems given by Beyth-Marom, Dekel, Gombo, and Shaked (in press).

The tutorial was given to 35 subjects in the usual large classroom groups and to 67 subjects who were run in small groups (4-7 people), with fewer other tasks and with small battery-powered calculators available for use.

For all four groups, none of the subjects knew when they completed the first form that they would later be given the second, Control form (one Tutorial subject asked the experimenter whether she was supposed to remember it all and was told no).

## Results

Lightbulb vs. Dyslexia. In order to compare the answers given to the two different problems, we counted the number of correct answers (for this count we required two-digit accuracy) and also tallied the number of answers for each problem in seven categories:

1. Too Low: Answers falling more than .10 below the correct answer. For the Lightbulb problem, this range was .00-.30; for Dyslexia, .00-.17.

2. About Right: Answers that were within .10 of the correct answer, including all correct answers. For the lightbulb problem, this range was .31-.51; for Dyslexia, .18-.38.

3. Middling: Answers greater than .10 above the correct answer but below the diagnosticity (Lightbulb, .39-.94; Dyslexia, .52-.79).

4. Diagnostic: Answers that were equal to the diagnosticity value stated in the problem (Lightbulb, .80; Dyslexia, .95).

5. Way High: Answers greater than the diagnosticity but not exceeding 1.00.

6. Outside: Negative answers and answers greater than 1.00.

7. None: No numerical answer given.

A comparison of the distributions of responses between the two problems showed that the problem with the larger range for a given category had more reponses in that category. For example, across all groups, 34% of the responses fell Too Low for the Lightbulb problem but only 20% were Too Low for Dyslexia. For the Dyslexia problem, 19% of all responses were Middling whereas only 6% were Middling for Lightbulb. However, the response categories of special interest had equal ranges across the two problems, and for these, About Right, Diagnostic, Outside, and None, the distributions for the two problems were remarkably similar. Thus we collapsed the data across the two problems.

Large vs. small groups. The tutorial condition was given in both large group and small group administration. The distributions of responses in the seven categories, collapsed across problems, did not differ for the two administrations. Indeed, exactly the same percentage of subjects gave the right answer. We thus collapsed the data across this variable, except for the error analyses to be discussed later.

Main results. The primary results of the experiment, the proportion of subjects giving answers in each category for each group, are shown in Table 3. The percentage of exactly correct responses are shown in parentheses because these percentages are included in the About Right category. The first column gives

results for the Control group for both administrations; thus it is based on two responses from each of 41 subjects. The columns labeled "2nd" show the responses to the second administration for the other three conditions; this was always the Control form.

---

Insert Table 3 about here

---

The results show that the List condition had no effect. Both the first (List) and second (Control) administrations showed results highly similar to the Control group, which, in turn, had results similar to previous experiments (e.g., Bar-Hillel, 1980). In contrast, the Algorithm and Tutorial conditions showed striking effects; no subjects gave a reponse equal to the diagnosticity value and about 40% gave responses close to the correct response. For the Algorithm group, this improvement did not generalize to the second, Control, problem; that distribution looks like the Control distribution. One might suppose that if the algorithm would have any generalizable effect, that effect might be limited to the 29 subjects who arrived at about the right answer when using it. However, when presented with the second, control problem, 16 of these 29 subjects (55%) responded with the diagnosticity and only one gave about the right answer.

The Tutorial group did appear to learn something. When they were given the Control problem, 31% gave about the right answer whereas only 9% gave the diagnosticitity value. On this second

problem 23% were able to come up with the correct answer accurate to two decimal places, although they had no guidance in front of them for doing so.

Is your answer sensible? After completing the first problem, the subjects in the Algorithm and Tutorial groups were asked whether the answer arrived at seemed sensible to them. The answers to this question are shown in Table 4. For both groups, the majority of subjects who answered the question said yes. For neither group was the proportion of Yes answers significantly different for those whose answer was about right than for the other subjects.

---

Insert Table 4 about here

---

The Algorithm group were then asked, "If you answered No, what do you think is a sensible answer?" Only 20 of the 26 "No" subjects gave a revised answer. Of these revised answers, only one was close to correct; this subject had perfectly performed the algorithm, arriving at an answer of .41 to the Lightbulb problem, but said that a sensible answer was .35. Eight subjects gave the base rate, six subjects gave the diagnosticity, and there were five other responses. In all, 12 of the 20 revised responses were in the Too Low range, supporting the finding shown in Table 4 that most of the Algorithm subjects who had originally calculated a low number found it sensible.

Errors. The subjects in the Algorithm and Tutorial tasks made numerous errors. We categorized and tallied all errors. Table 5 shows the errors of the Algorithm group. The results are shown separately for those who received the Lightbulb problem and those who received the Dyslexia problem because we found significant differences between the two groups. The Dyslexia group were particularly prone to the error of taking the wrong information from the story. The Dyslexia story differs from the Lightbulb story (and from stories used in previous research) by expressing the two pieces of diagnostic information in two different ways:

> . . .For children who really have dyslexia, the screening test is positive (indicating dyslexia) 95% of the time. But it also gives a positive (dyslexia) result for 5% of the normal children, the ones who do not have dyslexia.

Many subjects were apparently confused by this wording, so that when the algorithm asked, "What percentage of the time is the screening test able to correctly identify children that actually do not have dyslexia?", 45% of the subjects filled in 5%. One subject went so far as to write us a note in the margin saying that this question was incorrectly worded.

---

Insert Table 5 about here

---

The algorithm several times required subjects to copy a previous calculation into a new spot. About a third of the subjects made errors in following these simple directions.

We categorized arithmetic errors as (a) errors in sign, (b) addition or subtraction, and (c) multiplication or division. Within the multiplication or division errors we futher distiguished decimal errors, upside-down division, and other errors. Upside-down division is the calculation of the inverse of the indicated division; for example:

$$50 \div 2 = .04 \quad \text{or} \quad 2 \div 50 = 25 \ .$$

The Dyslexia subjects showed a significantly greater frequency of one or more arithmetic errors, $\chi^2 = 7.03$, $p < .01$, a finding we cannot explain.

An incomplete algorithm received only one tally for that reason, regardless of how many steps were omitted.

The errors made by the Algorithm subjects sometimes led to absurd answers; as shown in Table 3, 22% of the answers were outside the range of permissible probabilities, either negative or greater than 1.00. The largest answer was 4934.4; this subject made three decimal errors, one copying error, one multiplication error, one sign error, and ended with an upside-down division. The answer was judged not to be sensible; the sensible answer given was 2%.

Overall, Table 5 shows a discouraging inability of our subjects, the vast majority of whom are college students, to follow

simple instructions and perform simple arithmetic operations. Table 6 tells a similar story about the Tutorial group.

---

Insert Table 6 about here

---

The coding of errors in Table 6 is different from that used in Table 5 because we were interested primarily in conceptual errors. Thus the errors are organized according to the steps we taught the subjects to use to solve the problem.

Labeling the rows and columns was difficult for many subjects. Our code of "Confusing" means that the labeling confused the coder (the first author); it may not have confused the subject. "Wrong" labels included, for example, labeling one of the rows "Don't have dyslexia but test shows they do" or labeling the columns "Dyslexia children: test used/test not used" and the rows "Non-dyslexia children: screening test used/test not used."

In Table 6 the subcategory "Bollixed" means that the coder was unable to determine (or imagine) where the numbers came from.

Most Tutorial subjects were able to start with 1000 as a total and correctly allocate that according to the base-rate information given in the problem. Previous attempts to write a tutorial led to many errors in identifying the base rate; the present tutorial strongly emphasized this aspect, apparently with reasonable success. In contrast, our subjects had great difficulty in allocating numbers to the four cells. The most common error was to

put the numbers in the wrong cells, specifically (as exemplified by the Dyslexia problem):

| | | First Graders: Have Dys. | First Graders: No Dys. | |
|---|---|---|---|---|
| Test says: | Have Dys. | 19 | 931 | |
| | No Dys. | 1 | 49 | |
| | | 20 | 980 | 1000 |

The tutorial did not warn about this particular error. When no other errors are made, this error yields an answer equal to the base rate. Base-rate answers were given by 17% of the Tutorial subjects.

Although both the subjects given the Tutorial in a large-group setting and those given it in a small-group setting produced the same percentage of correct answers (31%), there was a significant difference, $\chi^2 = 7.37$, $p < .01$, in the proportion of subjects who made one or more arithmetic errors, 43% for the large groups but only 18% for the small groups, for whom hand-held calculators were available. Otherwise, the distributions of errors did not differ for the two types of administration.

Discussion

Base-rate problems are difficult problems. Most college students cannot do them correctly without substantial help. Indeed, Eddy (1982) has shown that the authors of authoritative medical texts, who presumably have much more education and

sophistication than college students, frequently make errors in understanding the significance of base rates in interpreting mammograms (tests for breast cancer).

Our least potent aid, asking subjects to list relevant factors, was entirely ineffective. This result is consistent with the view that subjects do not have the knowledge necessary to solve base-rate problems. Thus, thinking harder about the problem doesn't help.

The algorithm improved performance only when it was in front of the subjects; it had no effect on the second, unaided problem. Our instructions did not suggest that the subjects should study the algorithm or try to see what process it represented. Apparently, the subjects got caught up in putting the right numbers in the right places without gaining any insight into the problem or its solution.

The effective aid was a specially-written tutorial on how to solve base-rate problems. When the tutorial was in front of them, 42% of these subjects arrived at about the right answer. Moreover, when presented with the second, control problem, 31% gave about the right answer and only 9% gave the diagnosticity.

There were two main barriers to success in the tutorial condition. First, the tutorial appears, in retrospect, to have given insufficient attention to the task of allocating numbers to cells. This conceptual problem might be rectified by re-writing and expanding the tutorial.

Second, the subjects' elementary arithmetic skills were weak. When no calculator was made available, 43% of our Tutorial subjects made one or more arithmetic errors. For the Algorithm subjects, who were not given the minimal arithmetic examples contained in the tutorial, the arithmetic error rate was 51%. One would not expect to overcome this difficulty easily. We wonder how those people balance their check books or assess bargain prices. Presumably, they just don't.

Nonetheless, the tutorial approach holds great promise. Although it appears that most college students do not start with the knowledge required to solve base-rate problems, they can be taught it successfully in a relatively short period of time (about half an hour) without individual tutoring, practise, or feedback.

Two further problems remain. First, people who are taught to perform well on base-rate problems may not be able to discriminate between base-rate problems, in which their new training is relevant, and other, somewhat similar problems that cannot be solved using this approach, as the results of Fischhoff and Bar-Hillel (1984) suggest. Second, those trained in the laboratory on story problems may not be able to recognize base-rate problems that arise elsewhere.

If a tutorial could be written that solved the first problem--when not to use the technique--it might form the basis for a larger educational program to address the second problem--recognizing base rates in daily life. We share the optimism of Nisbett, Kranz,

Jepson, and Kunda (1983), who suggested that "training in statistics should promote statistical reasoning even about mundane events of everyday life. . ." (p. 347).

References

Ajzen, I. (1977). Intuitive theories of events and the effects of base rate information on prediction. Journal of Personality and Social Psychology, 35, 303-314.

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

Bar-Hillel, M. (1983). The base rate fallacy controversy. In R. W. Scholz (Ed.), Decision making under uncertainty (pp. 39-61). Amsterdam: Elsevier.

Beyth-Marom, R., & Arkes, H. R. (1983). Being accurate but not necessarily Bayesian: Comment on Christensen-Szalanski and Beach. Organizational Behavior and Human Performance, 31, 255-257.

Beyth-Marom, R., Dekel, S., Gombo, R., & Shaked, M. (In press). An elementary aproach to thinking under uncertainty. Hillsdale, NJ: Lawrence Erlbaum.

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

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

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

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

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

MacGregor, D., Lichtenstein, S., & Slovic, P. (1984). Structuring knowledge retrieval. Decision Research Report 84-14.

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

Shaugnessy, J. M. (1983). The psychology of inference and the teaching of probability and statistics: Two sides of the same coin? In R. W. Scholz (Ed.), Decision making under uncertainty (pp. 325-350). Amsterdam: Elsevier.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), Progress in social psychology (pp. 49-72). Hillsdale, NJ: Erlbaum.

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

## Author Notes

This research was supported by the Army Research Institute under contract MDA 903-83-C-0198.

We thank Paul Slovic for his helpful comments.

Requests for reprints should be sent to Sarah Lichtenstein, Decision Research, 1201 Oak Street, Eugene, OR 97401.

Table 1

The Base Rate Problems

| Light Bulb |
| --- |

Consider the following problem:

A light bulb factory uses a scanning device which is supposed to put a mark on each defective bulb it spots in the assembly line. Eighty-five percent (85%) of the light bulbs on the line are OK; the remaining 15% are defective.

The scanning device is known to be accurate in 80% of the decisions, regardless of whether the bulb is actually OK or actually defective. That is, when a bulb is good, the scanner correctly identifies it as good 80% of the time. When a bulb is defective, the scanner correctly marks it as defective 80% of the time.

Suppose someone selects one of the light bulbs from the line at random and gives it to the scanner. The scanner marks this bulb as defective.

What is the probability that this bulb is really defective?

(table continues)

Table 1 (continued)

| Dyslexia |
| --- |
| Dyslexia is a disorder characterized by an impaired ability to read. Two percent (2%) of all first graders have dyslexia. A screening test for dyslexia has recently been devised that can be used with first graders. The screening test is cheap and easy to administer; it identifies those children who will later be given a more extensive test to determine for sure whether the child has dyslexia. The screening test is not completely accurate. For children who really have dyslexia, the screening test is positive (indicating dyslexia) 95% of the time. But it also gives a positive (dyslexia) result for 5% of the normal children, the ones who do not have dyslexia.<br><br>A first grader is given the screening test and the result is positive, indicating dyslexia.<br><br>What is the probability that the child really has dyslexia? |

Table 2

Algorithm for the Lightbulb Problem

(A) Out of 1,000 light bulbs produced by the factory, how many are defective? Multiply the percentage of defective bulbs by 1,000. (First convert the percentage value to a decimal value before multiplying.)

1,000 x _____ = _____(A)

Proportion of Defective Bulbs

(B) Subtract your estimate in (A) from 1,000 to get the number of bulbs out of 1,000 that are NOT defective.

1,000 - (A)_____ = _____(B)

(C) What percentage of the time is the scanner able to correctly identify light bulbs that are actually defective? (from the problem) _____(C)

(D) What percentage of the time is the scanner able to correctly identify light bulbs that are actually not defective? (from the problem) _____(D)

(table continues)

Table 2 (continued)

(E) Look over the following table:

LIGHT BULBS ARE:

| | Actually Defective | Not defective | |
|---|---|---|---|
| Scanner Says IS Defective | Box # 1 | Box # 4 | _____ (L) |
| Scanner Says IS NOT Defective | Box # 2 | Box # 3 | |
| | _____ (A) | + _____ (B) | |

(F) Write the number of defective light bulbs from (A) on the line labeled (A) in the table above, just below Box #2.

(G) Write the number of non-defective light bulbs from (B) on the line labeled (B) in the table above, just below Box #3.

(H) Multiply the percentage value in (C) by your estimate from (A). (First convert the percentage value to a decimal value before multiplying.)

(A) _____ x (C) _____ = _____(H)

Write your value for (H) in Box #1.

(table continues)

Table 2 (continued)

(I) Subtract your value in (H) from your value in (A).

(A )_____ - (H) _____ = _____(I)

Write your value for (I) in Box #2.

(J) Multiply the percentage value in (D) by your estimate from (B). (First convert the percentage value to a decimal value before multiplying.)

(B) _____ x (D) _____ = _____(J)

Write your value for (J) in Box #3.

(K) Subtract your value in (J) from your value in (B).

(B) _____ - (J) _____ = _____(K)

Write your value for (K) in Box #4.

(L) Add the numbers in Boxes #1 and #4.

Box #1_____ + Box #4_____ = _____(L)

Write your value for (L) on the line labeled (L), to the right of the boxes.

(M) To get the final answer, divide your value in Box #1 by your value for (L).

Box #1_____ ÷ (L) _____ = _____(M)

Table 3

Distributions of Answers, in Percentages, for All Groups

| | Control | List | | Algorithm | | Tutorial | |
|---|---|---|---|---|---|---|---|
| | Both | 1st | 2nd | 1st | 2nd | 1st | 2nd |
| Too Low | 21 | 27 | 20 | 24 | 30 | 30 | 35 |
| About Right | 9 | 7 | 5 | 38 | 3 | 42 | 31 |
| (Exact) | (4) | (1) | (0) | (22) | (0) | (31) | (23) |
| Middling | 15 | 14 | 23 | 7 | 11 | 6 | 12 |
| Diagnostic | 48 | 43 | 45 | 0 | 51 | 0 | 9 |
| Way High | 2 | 5 | 6 | 5 | 3 | 7 | 4 |
| Outside | 0 | 0 | 0 | 22 | 1 | 4 | 4 |
| None | 5 | 5 | 1 | 4 | 1 | 11 | 5 |
| No. of Ss | 41[a] | 86 | | 76 | | 102 | |

[a]Each of 41 subjects contributed two responses to this distribution.

Table 4

Frequencies of Answers to the Sensibleness Question

| | Algorithm | | | Tutorial | | |
|---|---|---|---|---|---|---|
| | Yes | No | % Yes | Yes | No | % Yes |
| Too Low | 14 | 3 | 82 | 20 | 11 | 61 |
| About Right | 16 | 13 | 55 | 36 | 6 | 86 |
| Too High | 6 | 3 | 67 | 9 | 4 | 69 |
| Outside | 10 | 7 | 59 | 3 | 0 | 100 |
| Total | 46 | 26 | 64 | 68 | 21 | 76 |
| Not answered | | 4 | | | 13 | |

Table 5

Error Analysis for the Algorithm Group, in Percentages

| | All (n=76) | Lightbulb (n=29) | Dyslexia (n=47) |
|---|---|---|---|
| Wrong Info from Story | 47 | 28 | 60 |
| One or More Copying Error | 32 | 28 | 34 |
| One or More Arithmetic Error | 54 | 34 | 66 |
| Sign Error | 8 | 7 | 9 |
| Addition or Subtraction | 13 | 14 | 13 |
| Multiplication or Division | 51 | 31 | 64 |
| Decimal Error | 22 | 17 | 26 |
| Upside-down Division | 13 | 3 | 19 |
| Other | 30 | 21 | 36 |
| Incomplete Algorithm | 4 | 3 | 4 |
| No Errors | 22 | 41 | 11 |
| Mean No. Errors per Subject | | 2.21 | 2.89 |
| Most Errors by One Subject | | 14 | 12 |

Table 6

Error Analysis for the Tutorial Group, in Percentages

| | |
|---|---|
| Labeling: | |
| No error | 63 |
| Confusing | 22 |
| Wrong | 15 |
| Allocating the base rate[a]: | |
| No error | 84 |
| Used wrong rate | 10 |
| Arithmetic error | 3 |
| Bollixed | 5 |
| Incomplete | 2 |
| Allocation to cells[a]: | |
| No error | 37 |
| Used wrong rate | 8 |
| Put in wrong cells | 38 |
| Arithmetic error | 11 |
| Bollixed | 12 |
| Incomplete | 5 |
| Crossing out irrelevant: | |
| No error | 73 |
| Wrong | 15 |
| Incomplete | 12 |

(table continues)

Table 6 (continued)

| | |
|---|---|
| Final computation[a]: | |
| No error | 67 |
| Used wrong numbers | 4 |
| Arithmetic error[b] | 9 |
| Upside-down division | 7 |
| Bollixed | 7 |
| Incomplete | 11 |
| Overall, one or more arithmetic error | 26 |
| Overall, perfect or nearly perfect[c] | 31 |

[a]More than one error can occur in this category; %'s total more than 100. [b]Excluding upside-down division. [c]Errors only in labeling.

# Appendix

## Tutorial (Lightbulb version)

Consider the following problem:

A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

(a) 90% of the cabs in the city are Green and 10% are Blue.
(b) a witness identified the cab as Blue.

The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 70% of the time and failed 30% of the time.

What is the probability that the cab involved in the accident was Blue rather than Green?

Research has shown that people often have trouble answering problems like this. In this portion of today's experiment, we are presenting you with a mini-tutorial to see if instruction will help you solve such problems. Please read through the tutorial carefully. We have allowed time in the experiment for you to do that.

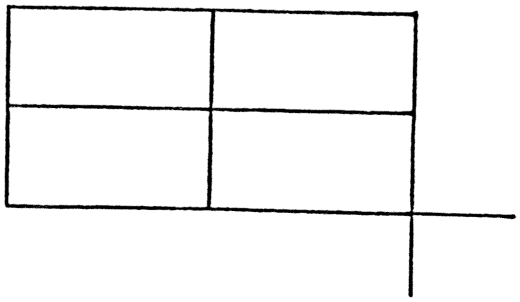
### Tutorial

The class of problems here addressed are problems for which two kinds of information are given and a probability is requested. One kind of information is about the population or populations in question. The other kind of information is specific to the case at hand.

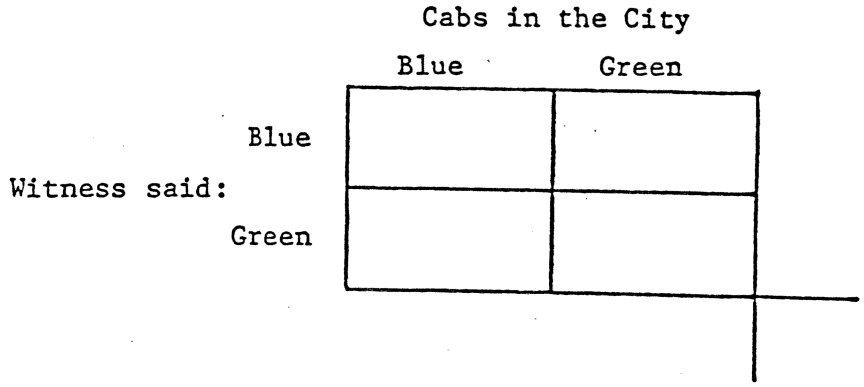
In the problem given above, the population is the population of cabs in the city. The population information is that 90% of the cabs are Green and 10% are Blue. The specific information concerns the specific cab that was involved in a hit and run accident. The witness said that that specific cab was Blue. But we also know about this testimony that the witness is not perfectly accurate. The witness is able to correctly identify the color of the cab 70% of the time.

The way most people usually go wrong in solving these problems is that they concentrate too much on the specific information and tend to neglect the population information. Maybe the specific information seems more immediately relevant to them. Or perhaps they just don't know how to go about combining the information to produce a single answer. Here is a way of doing just that:

Step 1. Draw a table. Begin by drawing a "two-by-two" table, that is, a diagram with two rows and two columns, like this:



Step 2. Label the table. We'll label the columns for the population information. The population is cabs in the city, which are either Blue or Green. The rows get the specific information, that is, the witness testimony, which was Blue--but for completeness, we'll also label the other row Green, because the witness could have said Green. So now our table looks like this:

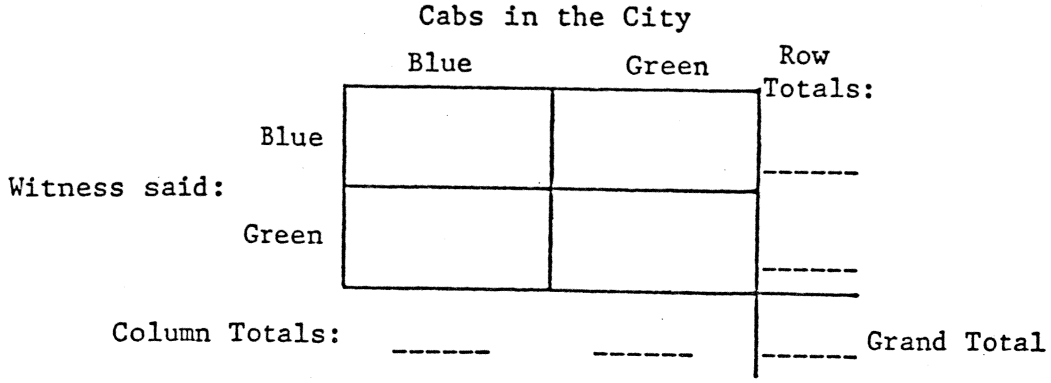



Labeling the table is not quite as simple as it may first appear. Notice that the sub-labels, "Blue" and "Green", are the same for the rows and the columns. This should generally be true in such problems. It would be a mistake to label the rows according to whether the witness was accurate or inaccurate:

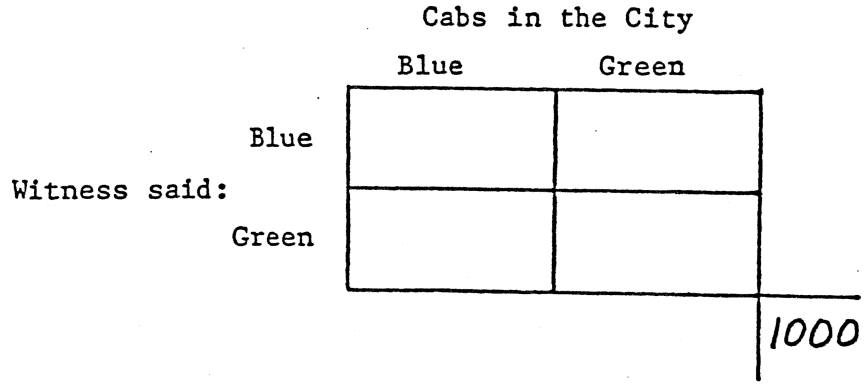
Witness was:
Right
Wrong

The problem could be solved with such labeling, but not using the method we are teaching you here. In general, the sub-labels are the two possible states of the world. The main labels (e.g., "Cabs in the City" and "Witness said:") indicate the source of information. One source is always population information (here, the relative number of cabs in the city); the other source is always specific information (here, what the witness said).

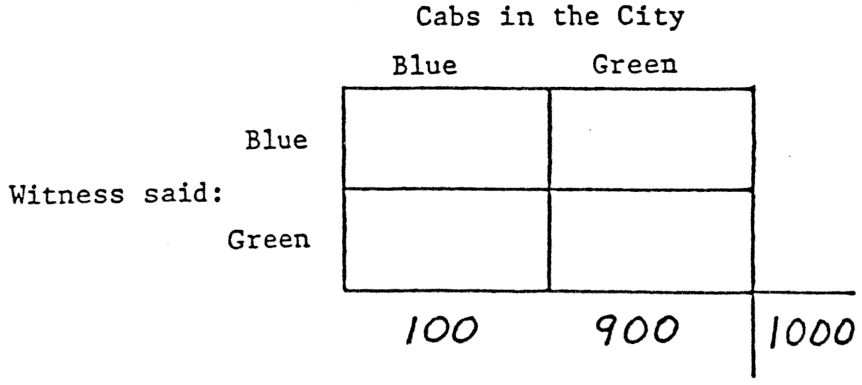
Notice that if there were numbers in the four cells of the table, we could calculate row totals and column totals and a grand total for the whole table. The places for these totals are shown below with dashed lines.




Step 3. Assign an arbitrary grand total. To get started, we'll fill in the grand total. That should be the total number of cabs in the city. But we don't know how many cabs there are in the city. So we pick an arbitrary total of 1,000. We could use 10 or 100 (or any other number), but using 1,000 will make later calculations easier.




Step 4. Estimate the population totals. If there were 1,000 cabs in the city, how many of them would be Blue? According to the story, 10% are Blue. That means 10 out of every 100 or 100 out of every 1,000 are Blue. That number, 100, is the left column total. The rest are Green. So 1,000 - 100 = 900 is the right column total. We put these column totals into the table:



WARNING. The method we're teaching you for solving these problems won't work if you start out estimating the wrong totals. It's important in this step to correctly identify which part of the problem gives population information and which gives specific information. The population information is general, background information that does not indicate any specific case. The specific information fingers a particular case.

Step 5. Fill in the cells. Working with each total, divide it among its two cells. First, for the 100 Blue cabs, how many would the witness correctly see as Blue, and how many would the witness incorrectly see as Green? The story states that the witness is correct 70% of the time. So:

$$\begin{array}{r} 100 \\ \times\ .70 \\ \hline 70 \end{array}$$

70 is the number of Blue cabs the witness would correctly call Blue, and the remaining, 100 - 70 = 30, are the number of Blue cabs the witness would incorrectly call Green.

Now consider the 900 Green cabs. Again the witness' accuracy is 70%:

$$\begin{array}{r} 900 \\ \times\ .70 \\ \hline 630 \end{array}$$

630 is the number of Green cabs the witness would have correctly called Green. This number, 630, goes in the Green-Green cell. The rest of the Green cabs, 900 - 630 = 270, is the number of Green cabs the witness would have incorrectly called Blue.

Our table now looks like this:

| | | Cabs in the City: Blue | Cabs in the City: Green | |
|---|---|---|---|---|
| Witness said: | Blue | 70 | 270 | |
| | Green | 30 | 630 | |
| | | 100 | 900 | 1000 |

Comment. Notice that we now could, if we wished, find the last two totals, the total number of times the witness would have said "Blue," rightly or wrongly:

$$70 + 270 = 340$$

and the total number of times the witness would have said "Green," rightly or wrongly:

$$30 + 630 = 660.$$

These totals are not intuitively obvious. The reason is that these totals are the total number of times the witness says "Green" and "Blue." What the witness says depends not only on the witness' accuracy but also on the relative proportions of Blue and Green cabs the subject

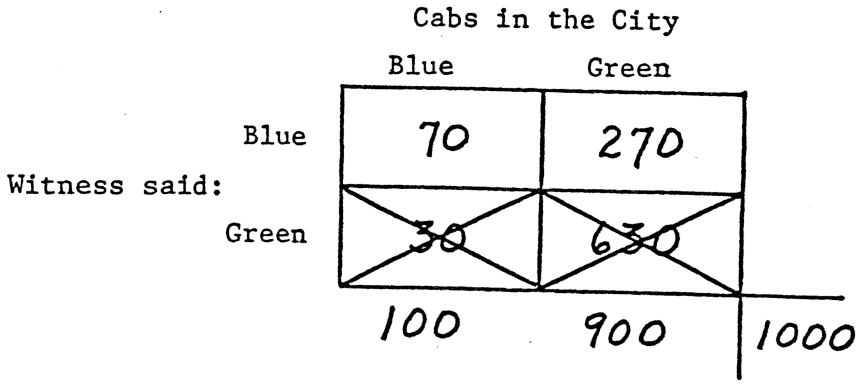
might have seen. You have to take both these facts into consideration to calculate the totals. In contrast, the population totals make a lot of sense, because they depend on only one kind of information, not two kinds. The total number of Blue cabs in the city is directly calculated as a percentage of the total number of cabs, regardless of what the witness might testify. This distinction is important because it shows you another way of telling, in any problem, which is the population information (that you start with in Step #4) and which is the specific information. The population information is information that directly translates into number totals. The specific information is information that does not translate into number totals because those number totals depend not only on the specific information but also on the population information.

In summary, here are two criteria (one discussed earlier) for telling which is which:

The population information:
(a) is general, background information and
(b) can be translated directly into number totals.

The specific information:
(a) specifies or identifies one case and
(b) cannot be directly translated into number totals because those totals also depend on the population information.

Step 6. Cross out the false. The witness in the story in fact testified that the cab was Blue. So the number of times the witness might have said "Green" is irrelevant to the problem. We cross out these false cells so we won't be tempted to use them in the next step:




Step 7. Find the needed probability. The two remaining cells are what we need to answer the question. They show that the witness would have said "Blue" correctly 70 times and would have said "Blue" incorrectly 270 times. From these two numbers we can get our probability.

If you're not used to thinking about probabilities, a nice way to think about them is to imagine that you fill an urn with 70 balls labeled "cab is really Blue" and 270 balls labeled "cab is really Green," for a total of 340 balls. Now sample one ball at random from

the urn. What is the probability that the ball will be labeled "cab is really Blue?" The answer is the number of "cab is really Blue" balls divided by the total number of balls in the urn:

$$\frac{70}{70+270} = \frac{70}{340} = .21$$ (well, it's really .2058...but we rounded it)

In other words, we divide the number in the TARGET cell by the sum of the two numbers left in our table. The TARGET cell is the one cell identified by both the specific information given in the problem ("a witness identified the cab as Blue") and the question asked at the end of the problem ("What is the probability that the cab involved in the accident was Blue?"). So the target cell is the "Cab is Blue/Witness said Blue" cell.

That's it. The answer, .21, is the probability that the hit-and-run cab was a Blue cab.

Are you surprised by the answer? Most people think that the correct answer should be .70, the same as the witness' accuracy. They tend to forget the population information, that is, they fail to notice that because there are so many more Green cabs than Blue cabs, there are also many more opportunities for the witness to be wrong when saying Blue.

Comment. While it's not necessary to solve the problem, it might help you to understand what's going on by thinking about this: What if the witness had testified that the cab was Green? Look back at the last table, the one with two crossed-out cells. Those crossed-out cells show 30 really Blue cabs and 630 really Green cabs. So the probability that the cab is really Green, if the witness said it was Green, is:

$$\frac{630}{630+30} = \frac{630}{660} = .95$$

This probability is higher than either the proportion of Green cabs in the city (90%) or the accuracy of the witness (70%). That's because in this case both pieces of information--the population proportion and the witness' testimony, point in the same direction, towards Green.

Intermediate probabilities like .21 are found only when the two pieces of information point in opposite directions: the witness said Blue but most cabs are Green.

That's the end of the tutorial. On the next page is a problem for you to do. Before doing the problem:

1. Review the tutorial to make sure you understand it.
2. Ask any questions you have.

When you are ready, proceed to the problem on the next page. We are interested in how effective the tutorial is in teaching you how to do such problems. So while you are doing the problem, feel free to:

1. Review the tutorial again.
2. Use a hand calculator.
3. Ask questions.

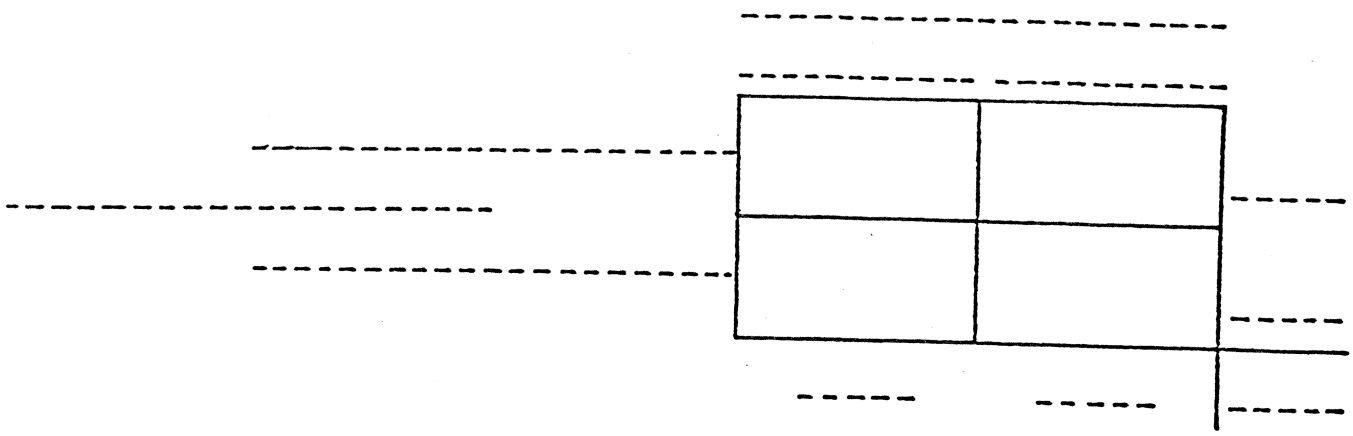
Please work the following problem using the method just described. We've drawn you a table to work with.

A light bulb factory uses a scanning device which is supposed to put a mark on each defective bulb it spots in the assembly line. Eighty-five percent (85%) of the light bulbs on the line are OK; the remaining 15% are defective.

The scanning device is known to be accurate in 80% of the decisions, regardless of whether the bulb is actually OK or actually defective. That is, when a bulb is good, the scanner correctly indentifies it as good 80% of the time. When a bulb is defective, the scanner correctly marks it as defective 80% of the time.

Suppose someone selects one of the light bulbs from the line at random and gives it to the scanner. The scanner marks this bulb as defective.

What is the probability that this bulb is really defective?



Step 1. Draw a table. Done.

Step 2. Label the table.

Step 3. Assign an arbitrary grand total. Use 1,000.

Step 4. Estimate the population totals. First decide which set of information is population information. Then divide the 1,000 into two parts, using information from the problem.

Step 5. Fill in the cells. Divide each of your estimated totals among its two cells, according to the information in the problem.

Step 6. Cross out the false. Cross out the two cells that are contradicted by the information given in the problem.

Step 7. Find the needed probability. Write the relevant numbers in the top and bottom of the fraction and convert the fraction to a decimal answer.

$$\frac{\#\text{ in target cell}}{\text{Sum of \#'s in both cells}} = \frac{\qquad}{\qquad} = \frac{\qquad}{\qquad} = \qquad , \text{ answer.}$$