Information selection and use in hypothesis testing: What is a good question, and what is a good answer?

LOUISA M. SLOWIACZEK

University at Albany, State University of New York, Albany, New York

JOSHUA KLAYMAN

University of Chicago, Chicago, Illinois

STEVEN J. SHERMAN

Indiana University, Bloomington, Indiana

and

RICHARD B. SKOV

Stevens Institute of Technology, Hoboken, New Jersey

The process of hypothesis testing entails both information selection (asking questions) and information use (drawing inferences from the answers to those questions). We demonstrate that although subjects may be sensitive to diagnosticity in choosing which questions to ask, they are insufficiently sensitive to the fact that different answers to the same question can have very different diagnosticities. This can lead subjects to overestimate or underestimate the information in the answers they receive. This phenomenon is demonstrated in two experiments using different kinds of inferences (category membership of individuals and composition of sampled populations). In combination with certain information-gathering tendencies, demonstrated in a third experiment, insensitivity to answer diagnosticity can contribute to a tendency toward preservation of the initial hypothesis. Results such as these illustrate the importance of viewing hypothesis-testing behavior as an interactive, multistage process that includes selecting questions, interpreting data, and drawing inferences.

Two of the critical elements of successful hypothesis testing are choosing a good question to ask and then knowing what to make of the answer. In a probabilistic environment especially, those goals are not always easily achieved. Suppose, for example, that you travel to the planet Vuma, where there are equal numbers of two kinds of creatures, Gloms and Fizos. You have information about the proportion of Gloms and of Fizos possessing a number of features, as illustrated in Table 1. For example, you know that 10% of Gloms wear hula hoops, as do 50% of Fizos. A creature approaches you, and you need to determine which kind of creature it is. The creatures are invisible,

but if you ask them whether or not they possess a particular feature, they will truthfully answer "yes" or "no." You have time for only one question. Given the features listed in Table 1, which feature would you ask about in order to best determine what kind of creature it is? How sure would you be about the creature's identity when you got a "yes" or a "no" answer?

Much recent work in hypothesis testing has focused on the information-gathering strategies subjects use. Results show that people are influenced by several different characteristics of questions when choosing what to ask, sometimes in contradiction to statistical prescriptions (Baron, Beattie, & Hershey, 1988; Evans, 1989; Klayman & Ha, 1987; Skov & Sherman, 1986). People show a clear preference for more diagnostic questions, that is, those that better distinguish the hypothesis from the alternative (Bassok & Trope, 1984; Devine, Hirt, & Gehrke, 1990; Trope & Bassok, 1982, 1983). Among the questions in Table 1, for example, subjects almost always ask about Features 1, 4, 7, or 8 (Skov & Sherman, 1986).

Sensitivity to the diagnosticity of questions is important because it leads to an efficient selection of tests. Efficient information gathering does not guarantee efficient or unbiased use of the information gathered, however.

This work was supported in part by Grant SES-8706101 from the Decision, Risk, and Management Science Program of the National Science Foundation to J. Klayman, who is affiliated with the Center for Decision Research, Graduate School of Business, University of Chicago. We thank Jonathan Baron, Jane Beattie, James Chumbley, Michael Doherty, Jackie Gnepp, Stephen Hoch, R. Scott Tindale, and two anonymous reviewers for helpful comments on earlier drafts. We also thank David Houston for his help with data collection and Marybeth Hamburger and Theresa Tritto for their help with data analysis. Requests for reprints should be sent to Louisa M. Slowiczek at the Department of Psychology, University at Albany, State University of New York, Albany, NY 12222.

Table 1
Proportion of Gloms and Fizos on Vuma
Possessing Each of Eight Features

	Gloms	Fizos
1.	10% wear hula hoops	50% wear hula hoops
2.	28% eat iron ore	32% eat iron ore
3.	68% have gills	72% have gills
4.	90% gurgle a lot	50% gurgle a lot
5.	72% play the harmonica	68% play the harmonica
6.	32% drink gasoline	28% drink gasoline
7.	50% smoke maple leaves	90% smoke maple leaves
8.	50% exhale fire	10% exhale fire

Whereas optimal information selection depends on the diagnosticity of the question, optimal revision of initial beliefs depends on the diagnosticity of the specific answers received. We propose that although people are fairly good at perceiving the diagnosticity of questions, they are prone to systematic misperceptions of the diagnosticity of the answers. Specifically, people are not sufficiently sensitive to the fact that different answers to any given question can carry very different amounts of information about the hypothesis and thus warrant different degrees of belief revision.

Diagnosticity of an Answer Versus Diagnosticity of the Question

According to Bayes' theorem, the diagnosticity of an answer or datum, D , depends on the likelihood ratio, $p(D|H_1)/p(D|H_0)$, in which H_1 is the hypothesis being tested and H_0 is its alternative. This likelihood ratio, multiplied by the prior odds of H_1 , yields the revised odds of the hypothesis in light of the new datum. Consider again the case of the visitor to Vuma, where there are equal numbers of Gloms and Fizos. The visitor might ask about Feature 4, "Do you gurgle a lot?" A "yes" answer provides evidence that the creature is a Glom; a "no" answer favors Fizo. However, these two potential data are not of equal value. A "yes" answer indicates a Glom only weakly; many "yes" answers also come from Fizos [$p(\text{Glom}|\text{yes}) = .64$, based on a likelihood ratio of 9/5 and prior odds of 1/1]. On the other hand, a "no" answer is strongly indicative of a Fizo; negative answers very rarely come from Gloms [$p(\text{Fizo}|\text{no}) = .83$, based on a likelihood ratio of 5/1].

Given that different answers can have different diagnosticities, how might one measure the overall diagnostic value of a question? Such a measure must take into account the chance of obtaining each possible datum (i.e., each possible answer) and the value of information each datum would provide if it occurred. One way of doing this is to compare the a priori likelihood of correctly accepting or rejecting the hypothesis with and without asking the question (Baron, 1985, chap. 4). For example, armed only with knowledge of the population base rates, a visitor to Vuma would have a .50 chance of correctly guessing the identity of an encountered creature. If the visitor were to ask about, say, Feature 4, the probability of correctly identifying the creature would increase. There is a .70 chance

that a randomly selected creature will answer "yes" to that question. If so, the visitor will guess Glom, and will have a .64 chance of being correct [$p(\text{Glom}|\text{yes}) = .64$]. There is also a .30 chance that the creature will answer "no." In that case, the visitor will guess Fizo and will have a .83 chance of being right [$p(\text{Fizo}|\text{no}) = .83$]. Thus, the a priori chance of guessing correctly after asking about Feature 4 is $(.70 \times .64) + (.30 \times .83) = .70$, an increase of .20 over the naive state. The expected increase in the probability of correct identification serves as a measure of the informativeness of the question. Another measure is the expected log likelihood ratio across possible answers (Klayman & Ha, 1987). For Feature 4, that would be $.70\log(9/5) + .30\log(5/1) = 0.388$.

Despite definitional complexities, the diagnosticity of a question is actually quite easy to estimate. With two initially equiprobable hypotheses, the expected increase in the chance of guessing correctly is proportional to the simple difference between the probability of a "yes" (or a "no") answer under H_1 and under H_0 (e.g., the difference between the two percentages shown for each feature in Table 1). The difference in probabilities is also highly correlated with the expected log likelihood measure.¹

We hypothesize that it is easier to perceive the diagnosticity of a question than it is the diagnosticity of each of its answers. A subject interested in the most diagnostic question can simply choose the one for which the two groups differ most in their likelihood of having the feature. On the other hand, the diagnosticity of individual answers to a question depends on the ratio of the groups' percentages for each answer rather than on the difference. People may continue to be influenced by the difference between percentages when judging the informativeness of different answers to a question, instead of attending to the different ratios appropriate for different answers. Because the principles of Bayesian statistics are not intuitively obvious to people (see Fischhoff & Beyth-Marom, 1983), people may assume that all answers to a good question will yield good information, without adjusting appropriately for the difference in diagnosticity of the different answers. Thus, in the above example, people will tend to overestimate the degree to which a "yes" answer indicates a Glom and/or underestimate the extent to which a "no" answer indicates a Fizo. We do not mean to imply that subjects will necessarily treat "yes" and "no" answers as exactly identical in all cases but that their judgments will be insufficiently sensitive to differences in answer diagnosticity.²

Data from two previous studies suggest such a pattern. Beach (1968) provided subjects with a series of problems involving two decks of cards with the letters A through F in different, known proportions in each deck. Cards were drawn from one, unidentified deck, and subjects, after seeing the letter drawn, indicated their subjective probabilities about which of the two decks had been selected. Subjects' probability judgments were affected by likelihood ratios, as they should have been, but were also correlated with the absolute magnitude of the constituent

probabilities. For example, drawing the letter E was seen as more diagnostic in choosing between decks with 39% and 13% E cards than in choosing between decks with 12% and 4% E cards. In both cases, the normatively relevant likelihood ratio is the same ($39/13 = 12/4$). Beach hypothesized that subjects regarded smaller probabilities as generally less informative, and their judgments were therefore correlated with the sum of the probabilities. However, a reanalysis shows that the mean responses in Beach's Figure 1 (p. 60) are better predicted by the difference between the probabilities (26% and 8%). In other words, when judging the diagnosticity of information, Beach's subjects were influenced by the differences in percentages, which corresponds to the diagnosticity of the question, and not just by the appropriate ratios. The same process, we propose, underlies insensitivity to answer diagnosticity. People underestimate differences in the diagnosticity of "yes" and "no" answers to the same question, because the difference in percentages is the same for both answers.

In another study, Birnbaum and Mellers (1983) asked subjects to estimate the likelihood that a given used car would last for 3 more years, given the opinion of an advisor. Subjects were given data on the advisor's prior history of correct and incorrect judgments, indicating (1) the advisor's ability in discriminating good used cars from bad and (2) the advisor's bias toward saying that cars were good or toward saying that they were bad. Ability to discriminate is analogous to question diagnosticity (i.e., the extent to which a positive opinion was more likely for good cars than for bad ones). The effect of bias is to introduce a difference in diagnosticity of different answers (i.e., for advisors with a positive bias, negative judgments were more diagnostic, and vice versa). In responding to positive or negative advice, subjects appeared to be more sensitive to differences in advisors' abilities to discriminate than to the effects of bias.

The findings of Birnbaum and Mellers (1983) and Beach (1968) suggest that people perceive the diagnosticity of a question more easily than the diagnosticity of specific answers. However, we know of no direct tests of the implication that people are insufficiently sensitive to differences in diagnosticity between alternative answers to the same question. Experiments 1 and 2 of the present paper are designed to test that hypothesis. In Experiments 1A and 1B, subjects were required to determine the category membership of individual items (creatures from a planet or balls from an urn). Insensitivity to answer diagnosticity implies that the subjects will have similar confidence in, say, the group membership of a creature who answers "yes" to a given question and one who answers "no," even when confidence should be quite different in those two cases. In Experiments 2A, 2B, and 2C, the subjects were tested for insensitivity to answer diagnosticity with the use of a different kind of judgment. In these studies, the subjects were asked to infer the composition of a population based on the answers provided by a random sample of members. Suppose, for example, one obtained a

random sample of 20 creatures, asked each independently about Feature 4 in Table 1, and received 14 "yes" answers and 6 "no" answers. Statistically, this is the expected result if the population is 50% Gloms and 50% Fizos. From such a population, one would expect the sample to have, on average, 10 Gloms, producing 9 "yes" answers and 1 "no" answer, and 10 Fizos, producing 5 "yes" answers and 5 "no" answers. Whereas each "yes" is modestly indicative of a Glom (9/5 odds = .64 probability), each "no" is strongly indicative of a Fizo (5/1 odds = .83 probability). Insufficient sensitivity to differential diagnosticity would lead a subject to conclude incorrectly that there are more Gloms than Fizos, because more responses were Glom-like than Fizo-like.

Consequences of Insensitivity to Answer Differences

As described above, insufficient sensitivity to the difference in diagnosticity of different answers to a question can lead subjects to over- or underestimate the information value of a datum. The net effect of this insensitivity depends on the questions asked. Features with symmetrical probabilities (e.g., present in 30% of one group and 70% of the other) present no difficulty: "yes" and "no" answers are in fact equally diagnostic, in opposite directions. With asymmetrical features, however, one answer is always more diagnostic than the other (e.g., Feature 4, discussed earlier, in which the feature is present in 90% of one group and 50% of the other). There is evidence that people prefer to ask about features with one extreme probability, compared with symmetrical questions of equal diagnosticity (see Baron et al., 1988, with regard to "certainty bias"). Thus, people may tend to select exactly those questions for which insensitivity to answer diagnosticity leads to inaccurate inferences.

If people are indeed prone to such inferential errors, what is the likely net effect on their beliefs? In particular, researchers have noted various tendencies that can lead to a bias toward hypothesis preservation (also known as confirmation bias or perseverance of beliefs), that is, a tendency toward unwarranted belief in the focal hypothesis over the alternative(s). Might insensitivity to answer diagnosticity contribute to hypothesis preservation? If one considers the full range of possible questions, the answer is no. People will generally overestimate the impact of less diagnostic answers, compared with their estimation of the impact of more diagnostic answers. For some questions, that will favor the focal hypotheses; for others, that will favor the alternative.

Research on information gathering demonstrates, however, that people's hypothesis testing is affected by which of a set of alternatives is favored as the working, or focal, hypothesis. Thus, a favored hypothesis and its alternatives may not receive equal treatment. For example, people have a clear preference for positive tests, that is, questions for which a "yes" answer favors the working hypothesis (see Evans, 1989; Klayman & Ha, 1987). Thus, subjects tend to ask about Features 4 and 8 in Ta-

ble 1 when asking "Is it a Glom?"; the presence of these features (a "yes" answer) favors the Glom hypothesis. Subjects asked to test "Fizo" will prefer questions about Features 1 and 7, since "yes" answers verify that hypothesis (Skov & Sherman, 1986). Still, insensitivity to answer diagnosticity would favor the focal hypothesis in some cases and favor the alternative in others. The person with a Glom hypothesis might choose Feature 4, in which the hypothesis-confirming "yes" answer is less diagnostic than the disconfirming "no," or Feature 8, in which the opposite is true.

Skov and Sherman (1986) found that subjects' preference for extreme probabilities was also hypothesis-dependent. Subjects tend to select questions about features that are either very likely or very unlikely under the working hypothesis, in preference to features with similarly extreme probabilities with regard to the alternative. According to this extremity preference, subjects testing the Glom hypothesis favor questions about Features 1 and 4, whereas subjects testing the Fizo hypothesis favor questions about Features 7 and 8, because these questions concern features that have extremely high (90%) or low (10%) probabilities given the focal hypothesis. (Subjects' overall preferences in information gathering reflected an additive combination of three favored qualities—diagnosticity, positive testing, and extremity.)

Skov and Sherman's (1986) findings suggest that people choose questions with extreme probabilities given the hypothesis, with less concern for extremity given the alternative. In most cases, then, the probability of the feature for the target group is more extreme than the probability for the nontarget group. This means that the hypothesis-confirming answer tends to be more common, and less diagnostic, than the disconfirming answer. For subjects with a Glom hypothesis, Feature 4 is an example, as discussed earlier. Feature 1 illustrates the effects of extremity preference in negative tests as well: This feature has an extremely low probability given the Glom hypothesis (10%) and a more moderate probability given Fizo (50%). Here, "no" answers confirm the hypothesis, and they are more frequent and less diagnostic. Indeed, the answer favoring the group with the more extreme probability is always less diagnostic,³ and that group is likely to be the target group of the hypothesis. If, as we propose, people are insufficiently sensitive to the difference in diagnosticity between different answers, the net effect will be that they will overestimate the impact of the hypothesis-confirming answers relative to hypothesis-disconfirming answers. On balance, then, inferential errors will tend to be in favor of the working hypothesis. Thus, the combination of a preference for extreme probabilities, given the hypothesis, and insensitivity to answer diagnosticity may contribute to hypothesis preservation.

In Experiment 3, we tested for the presence of two question-asking preferences that affect the consequences of insensitivity to answer diagnosticity. These are (1) a tendency to ask questions about features with extreme probabilities rather than more symmetrical features of

equal diagnosticity, and (2) a particular preference for questions in which the probability of the feature is extreme under the focal hypothesis.

In summary, the present study seeks to demonstrate that people are not sufficiently sensitive to differences in the diagnostic value of different answers to a given question and that this lack of sensitivity can produce systematic errors in judgments about individuals and about populations (Experiments 1 and 2). We also test whether people have information-gathering preferences that can exaggerate the effects of insensitivity to answer diagnosticity and bias those effects in the direction of hypothesis preservation (Experiment 3).

EXPERIMENT 1A

In this experiment (and the one that follows), we are interested in the subjects' confidence in inferences made from individual answers to questions. The cover story, adapted from Skov and Sherman (1986), involved a visit to the planet Vuma, which is inhabited by equal numbers of Gloms and Fizos. The subjects were provided with information about the percentages of Gloms and Fizos having certain features, using the four diagnostic features shown in Table 1 (Features 1, 4, 7, and 8). The subjects met several randomly selected creatures and were asked to determine the identity of each creature, based on the creature's "yes" or "no" response to one of these questions. The subjects predicted whether or not the creature was of a given type and then estimated the chances in 100 of the creature being that type. If subjects are sensitive to the differential diagnosticity of different answers to these questions, this sensitivity should be revealed in their probability estimates. For example, when the question concerns a feature present in 90% of Gloms and 50% of Fizos, a "yes" should ideally result in an estimate of a 64% chance of Glom, whereas a "no" should result in an estimate of 17% chance of Glom (i.e., an 83% chance of Fizo). On the other hand, insufficient sensitivity to answer diagnosticity implies that subjects' estimates will be significantly less different in these two cases than normative inferences would warrant.

Method

Subjects. One hundred sixty-eight subjects were recruited through the Indiana University student newspaper. The subjects were paid \$5 for their participation in the experiment. The subjects in this study did not participate in any of the other experiments reported in this paper. (The subjects in all subsequent experiments were similarly restricted from participating in more than one study.)

Procedure. The subjects were tested in groups of 20 or fewer. They received a booklet explaining that they would be visiting an imaginary planet on which two and only two types of invisible creatures lived. There were 1 million Gloms and 1 million Fizos on the planet. The subjects were told that (1) they would meet and talk to four randomly selected creatures, (2) an interpreter would ask one question of each creature (e.g., "Do you wear hula hoops?"), to which the creature would respond truthfully, and (3) the creature's answer would provide some information about its identity. In each case, the subjects were provided with the percentage of

Gloms and Fizos that had the relevant feature and the creature's "yes"/"no" answer. The features used were the ones shown as Features 1, 4, 7, and 8 in Table 1.

The subjects were divided into two conditions differing in focal hypothesis. In one condition, the subjects' task when they met a particular creature was to determine whether or not the creature was a Glom. In the other condition, the subjects were asked to determine whether or not the creature was a Fizo. Focal hypothesis was not expected to affect sensitivity to answer diagnosticity but was manipulated to test for possible interactions with other hypothesis-related variables such as positive versus negative tests and confirming versus disconfirming evidence.

The subjects met four creatures on the planet by turning to four separate pages in the booklet. On each page, the four features and corresponding percentages were listed, along with a reminder that 1 million Gloms and 1 million Fizos lived on the planet. Following the feature percentages, the question asked by the interpreter and the creature's "yes" or "no" answer were presented. Half the subjects were instructed to predict whether the creature was a Glom and to estimate the chances in 100 that it was a Glom. The other half of the subjects were asked to predict whether the creature was a Fizo and to estimate the chances in 100 that it was a Fizo. The subjects wrote their predictions and probabilities directly on the booklet page.

Each of the four randomly selected creatures that a subject met responded independently to a different one of the four features. The subjects received two "yes" and two "no" answers; half the subjects received the "yes" answers on Features 1 and 4, and half the subjects received the "yes" answers on Features 2 and 3. The order in which questions were asked was randomly determined for each subject.

Results

For purposes of analysis, probability estimates were recoded with respect to the hypothesis favored by the evidence. For example, the subjects might meet a creature that answers "yes" to the 90-50 feature. In the Fizo-hypothesis condition, a subject might indicate a 25% chance of the creature being a Fizo; this would be coded as a 75% chance of Glom. Responses in which the subject guessed a creature's identity in contradiction to the evidence were excluded from analysis.⁴ Analyses including focal hypothesis as a variable (i.e., instructions to predict whether the creature was a Glom vs. whether it was a Fizo) revealed no effect of this manipulation, as predicted; the data were therefore collapsed over focal hypothesis.

Several planned contrasts were tested by using the data summarized in Table 2. Each subject received one "yes" and one "no" answer of high diagnosticity and one "yes"

and one "no" answer of lower diagnosticity. The normative model predicts that there should be a strong effect of answer diagnosticity such that subjects should report probability estimates of .83 for highly diagnostic answers and .64 for less diagnostic ones, a difference of .19. Across subjects, the mean response was .72 for highly diagnostic answers and .66 for less diagnostic answers. This difference was significant [$t(164) = 4.63, p < .001$]. Thus, the subjects showed some sensitivity to the differential diagnosticity of answers, but this degree of sensitivity was quite small. The mean difference between high and low diagnosticity answers for each subject was .06, which was significantly less than the normative .19 [$t(164) = 9.72, p < .001$]. The mean response for highly diagnostic questions was significantly below the normative value of .83 [$t(164) = 9.82, p < .001$]. The mean for less diagnostic answers was marginally different from the normative .64 [$t(164) = 1.92, p < .07$].

It has been suggested elsewhere that subjects may be more influenced by "yes" answers than by "no" answers (see Sherman & Corty, 1984). We were able to check this in the present study: Each subject received two "yes" answers (one highly diagnostic, one less) and two "no" answers (also one of each diagnosticity). The mean response to "yes" answers was .71 and to "no" answers, .65. These were significantly different [$t(147) = 2.92, p < .01$], so there was some evidence of a tendency to see "yes" answers as more revealing than "no" answers. There was no evidence of an interaction between "yes"/"no" answers and answer diagnosticity.

It has also been suggested that people may give more weight to data that confirm the hypothesis being tested than they do to disconfirming data. Each subject in each hypothesis group received two answers that favored the hypothesis mentioned in the question and two answers that contradicted it. The design was balanced with respect to high and low diagnosticity, "yes" and "no" answers, and which of the two hypotheses was being tested. The mean response to confirming data was .67 and to disconfirming, .71. These were significantly different [$t(159) = 2.30, p < .03$], but not in the predicted direction.

In general, then, the subjects greatly underestimated the difference between moderately diagnostic and highly diagnostic answers. A normative difference of 19% in probability estimates was judged as a difference of only 6%. In particular, the subjects seemed to underestimate the diagnosticity of the more diagnostic answers (i.e., when the normative probability was .83). Individual-subject analyses confirm that most subjects did not consistently distinguish the high- and low-diagnosticity answers: Of 168 subjects, 83 (49%) at least once assigned a lower probability estimate to a high-diagnosticity answer than to a low-diagnosticity answer.

EXPERIMENT 1B

The subjects in Experiment 1A showed only small differences in estimated probabilities regarding categorical inferences drawn from answers that differed greatly in

Table 2
Normative Probabilities and Mean Estimated Probabilities
That the Creature Is From the More Likely Group
From Experiment 1A

Feature		Creature's Response			
		Yes		No	
%Glom	%Fizo	Normative Probability	Mean Estimate	Normative Probability	Mean Estimate
90	50	.64	.68	.83	.70
50	90	.64	.70	.83	.66
50	10	.83	.74	.64	.63
10	50	.83	.73	.64	.62

the strength of their implications. It is possible, though, that subjects would show a greater appreciation of the differences in information value between different possible outcomes if the task context were simpler. For example, one could use the urns-and-balls paradigm (alternatively portrayed as bookbags and poker chips) that was popular in the 1960s and 1970s for testing an understanding of probability concepts. A typical problem might describe two urns, Urn A and Urn B, each filled with a large number of balls of two colors. The proportion of each color in each of the urns is known. One of the two urns is selected at random, and the subject does not know whether it is Urn A or Urn B. A sample of balls is drawn from the chosen urn, with replacement. The subject is then asked to judge, based on the drawn sample, what the probability is that the chosen urn was Urn A (or Urn B).

We were unable to find any urns-and-balls study that directly tested sensitivity to the information-value difference between different answers (i.e., drawing different color balls), so we conducted two such tests. In these tests, we were concerned with the conclusions that subjects draw when different color balls are chosen. We again predicted that the subjects would be insufficiently sensitive to differences in diagnosticity.

Sixty-four subjects were tested in groups of 10 or fewer and received partial course credit toward an introductory psychology class for their participation. The subjects were told to imagine that they were blindfolded and brought into a room in which there were two urns filled with balls. Urn A has 90% red balls and 10% white balls; Urn B has 50% red balls and 50% white balls. A ball is randomly selected from one of the urns and it is a red (or a white) ball. Based on the color of the chosen ball, the subjects were asked to determine both the urn from which the ball was selected and the probability of the more likely urn. The subjects responded by writing their choice of urn and estimated probability in an experimental test booklet.

Thirty-two of the subjects were told that a red ball was drawn, and 32 were told that a white ball was drawn. Normatively, those receiving a red ball should judge a .64 probability of the 90/10 urn, and those receiving a white ball should judge a .83 probability of the 50/50 urn, reflecting the likelihood ratio of 9/5 for red and 5/1 for white. All but two of the 64 subjects selected the urn that was normatively more likely. The mean probability values for the 90/10 urn given the red ball and for the 50/50 urn given the white ball were both .77. These were, of course, not significantly different.

Our second test employed a within-subject design. This permitted a test of the subjects' judgments when their attention was drawn to the question of differences between the two answers (i.e., the color of ball chosen) by virtue of their being asked about both in succession. The procedures were identical to those described above, with the following exceptions. Fifty-six undergraduates were tested in small groups. All subjects received two problems. One problem was identical to the one used above; the other

was the same except that the colors were changed to green and blue, and the location was said to be "in another room." In one problem, the red ball (or the equivalent, green) was drawn; in the other, the white ball (or blue) was drawn. The order of color combinations and balls drawn was counterbalanced. The mean response for the 90/50 urn given the red or green ball (normative probability of .64) was .70; for the 50/50 urn given the white or blue ball (normative probability .83), it was .66. These did differ significantly [$t(55) = 2.36, p < .03$], but in the direction opposite to the normative prescription.

The results from these two urns-and-balls tests are consistent with the findings in the Glom and Fizo task. Indeed, there was no evidence of any sensitivity to answer diagnosticity when urns and balls were used. There was little differentiation even with a within-subject design that prompted the subjects to attend to possible differences in diagnosticity. What differentiation there was was, on average, in the wrong direction. (We are currently investigating possible reasons for this.)

EXPERIMENT 2A

The previous experiments tested the subjects' appreciation of answer diagnosticity in the context of judging category membership. In each case, the subjects were required to use the information given in order to judge the probability that the source of the datum (creature or urn) was of one categorical type or another. In the next set of experiments, we looked at insensitivity to answer diagnosticity in the context of a different kind of inference, namely inferring the makeup of a population from information received from a sample. We again used the story of a planet with two kinds of creatures, but with different circumstances.

In experiments 2A, 2B, and 2C, similar procedures were used. In each of these studies, the subjects were sent to a planet with a population of 100 creatures, comprising Gloms and Fizos in unknown proportions. Of these 100 creatures, 20 were sampled at random, and each was asked the same, single question. For half the subjects, that question concerned a feature present in 90% of Gloms and 50% of Fizos. For the other subjects, the feature was present in 50% of Gloms and 10% of Fizos. After receiving the (truthful) answers from the 20 creatures sampled, the subjects were asked to estimate, based on the answers received, how many of the planet's 100 creatures were Gloms and how many were Fizos. (Both groups were mentioned in the questions; focal hypothesis was not of interest in these studies.)

Table 3 shows two models for the normative best guess given different distributions of "yes" and "no" answers in the sample data. The maximum likelihood estimate is the number of Gloms in 100, G , that gives the highest probability of obtaining the observed sample of answers, S [i.e., maximum $p(S|G)$]. If one assumes that all possible populations, from 0 Gloms and 100 Fizos to 100 Gloms and 0 Fizos, are equally likely a priori, the maxi-

Table 3
Normative Values and Subjects' Estimates of the Number of Gloms in the Population of 100, Given Different Observed Samples of Creatures' Responses in Experiment 2A

Feature		Sample		Maximum Likelihood	Average Population	Subjects' Estimate
% Glom	% Fizo	Yes	No			
90	50	15	5	62.5	57.2	64.4
		14	6	50.0	47.5	64.6
		13	7	37.5	38.7	64.4
50	10	5	15	37.5	42.8	39.5
		6	14	50.0	52.5	41.9
		7	13	62.5	61.3	45.2

mum likelihood estimate is also the single most likely population distribution given the observed data [maximum $p(G|S)$]. The second model, the average population, is the expected average number of Gloms that would be observed if one aggregated across a very large number of planets with that data pattern. Specifically:

$$E(G) = \sum_{G=0}^{100} p(G|S) \cdot G.$$

The two normative estimates are not identical, but the differences are not important for our analyses.

As before, we hypothesize that subjects are insufficiently sensitive to the difference in information carried by different answers to the same question. Thus, subjects will give too much weight to the less diagnostic answers relative to the more diagnostic answers to a given question. For the 90-50 feature, "yes" answers are more likely but less diagnostic than "no" answers. We predict that subjects will overestimate the information carried by "yes" answers, which suggest the presence of Gloms, relative to "no" answers, which suggest Fizos, and will therefore overestimate the number of Gloms. Conversely, we predict that subjects will underestimate the number of Gloms when given information about the 50-10 feature, for which "yes" answers are more diagnostic and less frequent than "no" answers.

Method

Subjects. The subjects were 83 undergraduate students at Loyola University of Chicago. The subjects received partial course credit toward an introductory psychology class for their participation. We eliminated 10 subjects from the analyses because they either failed to complete the questionnaire or failed to follow instructions.

Procedure. The subjects were tested in groups of 10 or fewer. The subjects received a booklet that explained that they would be visiting a faraway planet. On this planet were a total of 100 creatures comprising two (and only two) kinds of creatures (Gloms and Fizos). Of the 100 creatures on the planet, the subjects would meet 20 randomly selected creatures. Forty of the subjects were told that 90% of Gloms and 50% of Fizos had a particular feature (the 90-50 condition), and 33 were told that 50% of Gloms and 10% of Fizos had the feature (the 50-10 condition). Based on a creature's answer to the question, the subjects were asked to respond in three ways. First, for each of the 20 creatures they met, the subjects predicted whether that creature was more likely a Glom or a Fizo. Second, the subjects estimated the chances in 100 that the creature was of the predicted type. Finally, after meeting all 20 creatures, the

subjects were asked to estimate how many of the 100 creatures on the planet were Gloms and how many were Fizos. The subjects responded by writing their predictions, probabilities, and base-rate estimates in an experimental booklet that contained the instructions as well as the particular question asked by each subject and the responses made by the sampled creatures.

Each condition was further divided into three groups that differed in the numbers of "yes" and "no" answers received from the 20 creatures and thus differed in the implied number of Gloms and Fizos among the planet's 100 creatures. The three distributions consisted of 5, 6, or 7 highly diagnostic answers out of 20. For the 90-50 condition, the highly diagnostic answer was "no," and for the 50-10 condition, the highly diagnostic answer was "yes." Within each distribution, the order of "yes" and "no" answers was randomly determined for each subject.

Results and Discussion

As in Experiment 1A, analyses of individual category judgments excluded responses in which the subjects' identification of the creature contradicted the evidence (13% of responses). No eliminations were necessary for analyses of population judgments.

Individual category inferences. First, consider the subjects' estimates of the probability that each individual was a Glom or a Fizo. These estimates were analyzed in an analysis of variance (ANOVA) with feature (90-50 or 50-10) and distribution (5, 6, or 7 highly diagnostic answers) as between-subject variables. In addition, there were two within-subject variables, creature's answer ("yes" or "no") and trial block (first 10 trials vs. last 10).

The results replicate the basic finding of Experiment 1, that subjects do not show much awareness that different answers to the same question carry different amounts of information about the identity of the source. Sensitivity to differences in diagnosticity would produce a feature (90-50 or 50-10) $\times$ answer ("yes" or "no") interaction, because "yes" is more diagnostic than "no" for the 90-50 feature, and vice versa for the 50-10 feature. This interaction was not significant ($F < 1$). The mean value for the high-diagnosticity features ("no" for the 90-50 feature and "yes" for the 50-10) was 68.9; the mean estimate for the low-diagnosticity features was 67.5. The mean estimates for "yes" and "no" answers were both 68.1. Thus, in contrast to Experiment 1A, there is no evidence for a tendency to give more weight to "yes" answers.

There were no main effects or two-way interactions involving distribution (5, 6, or 7 highly diagnostic answers) or block (first 10 trials vs. second 10). According to Bayesian statistics, these factors should affect not only the overall population estimate, as shown in Table 3, but judgments made about each individual. Specifically, the number of "yes" answers received at any point in the sequence provides information about the likely base rate of Gloms in the population, which should in turn affect the subjects' beliefs concerning the identity of the next creature. The nature of these sequence effects is rather subtle, however, and many studies have documented the difficulty people have with the use of statistical base-rate information (see, e.g., Bar-Hillel, 1980; Borgida &

Brekke, 1981; Sherman & Corty, 1984; Tversky & Kahneman, 1982a). Thus, it is not too surprising to find sequence and block effects missing in subjects' individual category inferences.

Inferences about the population. The second type of estimate required of the subjects came at the end of the 20-answer sequence, at which time they were asked to provide their best guess, based on the answers they had received, for the number of Gloms and Fizos (out of 100) on the planet. The last column of Table 3 shows the mean estimated number of Gloms.

Planned contrasts showed that, as predicted, the subjects' estimates were significantly higher than the maximum likelihood estimates for the 90-50 feature [$t(38) = 5.75, p < .001$] and significantly lower than maximum likelihood estimates for the 50-10 feature [$t(32) = 2.45, p < .03$]. Similar results occur when the average population estimate is used as the norm. It is particularly telling that 10 of 12 subjects guessed a majority of Gloms when the normative models indicate a clear majority of Fizos (the 13-7 sample), and 6 of 11 estimated a majority of Fizos when there should be a clear majority of Gloms (the 7-13 sample).

The results also show that, overall, the subjects overestimated the number of Gloms. For both normative estimates, the number of Gloms in the 90-50 group is equal to the number of Fizos in the 50-10 group. Thus, the normative overall mean estimated number of Gloms across all distributions is 50. An overall mean for estimates of Gloms that is higher than 50% would indicate that subjects give more weight to "yes" than to "no" answers. (For both the 90-50 and 50-10 cases, a "yes" answer is consistent with Glom.) The observed overall mean was marginally higher than 50% [$t(70) = 1.91, p < .06$]. This suggests that population estimates were somewhat more influenced by "yes" answers (which were evidence of Gloms) than by "no" answers (which were evidence of Fizos), even though the earlier individual-creature judgments showed no such bias.

Another interesting aspect of the observed estimates is how little they differ across the three different distributions for each feature. An ANOVA was performed on the subjects' population estimates with two between-subject variables, feature (90-50 or 50-10), and distribution (5, 6, or 7 highly diagnostic answers). There was a highly significant main effect of feature [$F(1,71) = 47.8, p < .001$], with means of 64.5 and 42.2 for the 90-50 and 50-10 features, respectively. The normative models predict no main effect of distribution (the means for 5-15 and 15-5, etc., are all .50), but there should be a strong feature $\times$ distribution interaction. Neither the main effect nor the interaction was significant (both $F_s < 1$).

This pattern of results can be modeled by a rather simple counting process. Assume that most creatures who answer "yes" are Gloms and most who answer "no" are Fizos. More specifically, assume proportion P of "yes" answers are Gloms and proportion Q of "no" answers are Fizos, based on the confidence with which each of

these classifications is made. If there are Y ("yes") answers in the sample of 20, the estimated number of Gloms in the sample is then $YP + (20 - Y)(1 - Q)$. This kind of simple summing process is much less reactive to changes in the number of "yes" answers in the sample than are the normative models. Insensitivity to answer diagnosticity can be represented by assuming that P and Q are the same for both the 90-50 feature and the 50-10 feature. The observed means of 64.5 for the 90-50 feature and 42.2 for the 50-10 feature fit a model that counts 81.2% of "yes" answers as Gloms and 74.5% of "no" answers as Fizos. The population estimates that result from this insensitive-counting model are (in the order shown in Table 3): 67.3, 64.5, 61.7, 39.4, 42.2, and 45.0.

EXPERIMENTS 2B AND 2C

The cover story and materials used in the following two experiments were the same as those used in Experiment 2A, but the procedure was varied. It is possible that the subjects were prompted to use a counting method in Experiment 2A by the requirement that they make a separate, sequential judgment about the identity of each creature. Also, the memory requirements of this sequential procedure may have precluded more sophisticated treatments of the aggregate data. Thus, in Experiment 2B, the subjects were given a list of the "yes" and "no" answers obtained from 20 randomly selected creatures; in Experiment 2C, the subjects were given the total number of "yes" answers and "no" answers received from the sampled creatures.

Methods

Experiment 2B was conducted with 133 subjects. The subjects were undergraduates at Indiana University. Instead of meeting each of the 20 creatures individually and predicting the category to which the creature belonged, the subjects were given a list of 20 answers obtained from 20 randomly selected creatures. The list consisted of a column with the appropriate number of Y s and N s, in random order. Experiment 2C was conducted with an additional sample of 94 subjects selected from introductory psychology courses at Indiana University. In this study, the subjects were simply given the summary total number of "yes" answers and "no" answers obtained from a random sample of 20 creatures. For both Experiments 2B and 2C, the subjects were tested in groups of 10 or fewer and received partial course credit for their participation. The subjects recorded their responses in the booklet that listed or summarized the "yes" and "no" answers provided by the sampled creatures.

Results and Discussion

The results for Experiments 2B (list format) and 2C (summary format) are shown in Table 4. With the list format, there was a strong main effect of feature [$F(1,127) = 14.1, p < .001$], with means of 59.0 and 48.2 for the 90-50 and 50-10 features, respectively. There was also a significant feature $\times$ distribution interaction [$F(2,127) = 3.27, p < .04$], with no main effect of distribution. Planned contrasts showed that estimates for the 90-50 feature were significantly above the maximum likelihood norms [$t(64) = 6.05, p < .01$], whereas estimates for

Table 4
Subjects' Estimates of the Number of Gloms in the Population of 100,
Given Different Observed Samples of Creatures' Responses
in Experiments 2B (List Format) and 2C (Summary Format)

Feature		Sample		Maximum Likelihood	Subjects' Estimates	
% Glom	% Fizo	Yes	No		Lists	Summaries
90	50	15	5	62.5	67.4	71.6
		14	6	50.0	56.4	66.9
		13	7	37.5	55.2	64.3
50	10	5	15	37.5	45.7	47.5
		6	14	50.0	47.6	41.6
		7	13	62.5	51.2	45.3

the 50-10 feature were not significantly different, on average, from the norm [$t(67) < 1.0$, n.s.]. Once again, looking across the 90-50 and 50-10 conditions, the subjects overestimated the number of Gloms [$t(131) = 2.77$, $p < .006$], indicating that subjects give greater weight to "yes" than to "no" answers.

With the summary format, there was a main effect of feature [$F(1,88) = 44.52$, $p < .001$], with means of 67.7 and 44.8. Neither the feature $\times$ distribution interaction nor the main effect of distribution was significant. Estimates for the 90-50 feature were significantly above the maximum likelihood estimate [$t(45) = 9.075$, $p < .001$], and estimates for the 50-10 feature did not differ significantly from the normative value [$t(47) = 1.55$, n.s.]. As in Experiments 2A and 2B, the subjects again overestimated the number of Gloms [$t(92) = 3.13$, $p < .002$].

The results with both the list format and the summary format substantially replicate the findings of the sequential procedure of Experiment 2A. In all three experiments, population estimates for the 90-50 feature were significantly higher than those dictated by normative models. Estimates for the 50-10 feature were in all cases lower than those for the 90-50 feature and less than or equal to the normative value. In all three experiments, the overall mean estimate was above 50. These results confirm the general pattern of insufficient sensitivity to differences in answer diagnosticity, as well as a general tendency to weight "yes" answers more than "no" answers. A basic counting model is a viable representation for the observed patterns, given that differences across distributions for a given feature were small.

EXPERIMENT 3A

The results of Experiments 1 and 2 suggest that when people evaluate the impact of evidence, they are insufficiently sensitive to the diagnosticity of the specific answer received. This produces systematic errors in inferences drawn from certain kinds of questions. What are the expected consequences of these inferential errors in the evaluation of a hypothesis? As we discussed in the introduction to this paper, the net effect of insensitivity depends on the kinds of questions people choose to ask. Based on the findings of Skov and Sherman (1986) and Baron et al. (1988), we hypothesize that (1) people prefer

questions with an extreme-probability feature over symmetrical questions of equal diagnosticity and (2) people particularly favor questions with an extreme probability given the focal hypothesis (i.e., questions for which the feature is either very common or very uncommon in the target group). If so, then people tend to select questions for which insensitivity to answer diagnosticity will produce inferential errors, and those errors will, on balance, tend to produce a bias in favor of the target hypothesis (see Note 3).

Experiments 3A and 3B were designed to test the hypothesized preferences for extremity of probabilities. These experiments differ from Experiments 1 and 2 because they examine how people choose questions to ask rather than how they respond to the answers to given questions. Experiment 3A is a replication of the Skov and Sherman (1986) study, using their materials and procedures. In Experiment 3B, following, information gathering was evaluated by using a different procedure and a different context, focusing on preferences for extreme-probability features in particular.

Method

Subjects. One hundred ninety-nine students were recruited from undergraduate psychology classes at Indiana University. The subjects received partial course credit for their participation.

Materials and Procedure. The subjects were tested in groups of fewer than 10 at a time. Each subject was given an experimental booklet that included instructions and an answer sheet. The instructions stated that they would be visiting a planet on which there were two, and only two, types of creatures, called Gloms and Fizos. There were 1 million of each type of creature on the planet. The subjects were told that they would meet a randomly selected creature and were to ask one question that would help determine the creature's identity. Ninety-eight subjects were told to determine whether or not the creature was a Glom; the other 101 subjects were told to determine whether or not it was a Fizo. The subjects were provided with the information shown in Table 1, and could choose to ask about any one of those eight features. The subjects responded by writing the question they selected on the answer sheet.

Results and Discussion

As expected, the subjects showed a very strong preference for the four diagnostic questions (1, 4, 7, and 8 in Table 1). These accounted for 196 of 199 choices. Table 5 shows the pattern of choices among the four diagnostic questions. Question types were defined relative to

Table 5
Choices of Best Question in Experiment 3A by Question Type

	Extreme $p(f H_1)$	Moderate $p(f H_1)$	Total
Positive	72	52	124
Negative	47	25	72
Total	119	77	196

the subject's focal hypothesis. *Positive* questions are those for which a "yes" answer favors the hypothesis. These were Features 4 and 8 for the Glom-hypothesis group, 1 and 7 for the Fizo-hypothesis group. *Extreme* $p(f|H_1)$ questions were those for which the probability of the feature given the focal hypothesis was either extremely high (.90) or extremely low (.10). These were Features 1 and 4 for the Glom-hypothesis group and Features 7 and 8 for the Fizo-hypothesis group. Note that in this choice set, questions with moderate probabilities given the focal hypothesis always had extreme probabilities given the alternative. There were no significant differences between the two focal-hypothesis groups; Table 5 shows results for both groups combined.

The results of this study replicated those of Skov and Sherman (1986). The subjects showed a significant preference for questions with extreme probabilities given the focal hypothesis [$\chi^2(1) = 9.0, p < .005$]. There was also a significant preference for positive questions [$\chi^2(1) = 13.8, p < .001$], with no significant interaction between extremity and positivity.

EXPERIMENT 3B

This study extends our tests of preferences for extremity in two ways. First, we use direct pairwise choice to examine preferences among questions with extreme probabilities given the hypothesis, given the alternative, and neither. Second, we introduce a different context (involving medical diagnosis) in addition to the extraterrestrial context used previously.

Method

Subjects. One hundred thirty-one students were recruited from undergraduate psychology classes at Indiana University. The stu-

dents received partial course credit for their participation in the experiment.

Materials and Procedure. Two sets of stimulus materials were created for use in the experiment. The first set included instructions asking the subjects to imagine that they were visiting a far-away planet on which lived equal numbers of two kinds of creatures (Gloms and Fizos). Each subject met 10 randomly selected creatures, 1 at a time. For each, the subject had to determine whether the creature was a Glom by asking one of two possible questions. Focal hypothesis was not varied in this experiment: All questions were expressed in terms of Gloms.

On each trial, the subject turned to a new page in an experimental booklet indicating two features and the percentages of Gloms and Fizos having each feature. The subjects asked a question based on the feature that they felt would best aid in determining whether the given creature was a Glom by placing a check mark next to the chosen feature on the booklet page. (No feedback was given concerning the creatures' answers or identities.)

The second set of stimulus materials included instructions that asked the subjects to imagine that they were physicians seeing patients with "acute splenosis," which could be one of two equally prevalent types, Glomitic or Fizotic. Each subject met 10 patients and had to determine whether the patient's splenosis was Glomitic by selecting one of two medical tests to perform. On each trial, the subjects turned to a new page of the experimental booklet that showed the feature associated with each of two medical tests and the percentage of Glomitic patients and Fizotic patients that had each feature. The subjects selected the medical test that they felt would best aid in determining whether the patient's splenosis was Glomitic by putting a check mark next to the chosen medical test on the booklet page.

The subjects were tested in groups of 20 or fewer. Seventy-one subjects received the planetary instructions, and 60 received the medical instructions. In both conditions, the subjects made 10 two-alternative forced choices in favor of the feature that would best aid in determining the creature or splenosis type. The two conditions used feature pairs that were identical with respect to the percentages of each group having each feature (see Table 6). The planetary condition used whimsical features similar to those used in Experiment 3A. The medical condition used fictitious medical-sounding features such as "excess dibumase in urine" and "a deficiency of salinate in blood." The order of the features within each pair was counterbalanced, and the 10 pairs were presented in random order.

Analyses of extremity preferences were performed by using the six feature pairs shown in Table 6. Within each pair, the two questions were of equal diagnosticity (as measured by the difference between the two groups' probabilities) and were either both positive tests or both negative. (The four unanalyzed pairs contrasted features of unequal diagnosticity or positivity.)

Table 6
Number of Subjects Choosing Each Feature in Pairwise Comparisons in Experiment 3B

Contrast	Feature Pair	Choices	
		Planetary	Medical
Extreme $p(f H_1)$ /No extremity			
1. Positive tests	90-55/65-30	44/27*	36/24
2. Negative tests	10-45/30-65	45/14*†	28/32
Extreme $p(f H_1)$ /Extreme $p(f H_0)$			
3. Positive tests	90-55/45-10	35/35‡	41/19*
4. Negative tests	10-45/55-90	44/27*	28/32
Extreme $p(f H_0)$ /No extremity			
5. Positive tests	45-10/65-30	48/23	29/31
6. Negative tests	55-90/30-65	35/36	21/9*†

*Significant preference, $\chi^2(1) > 3.8, p < .05$. †This question was inadvertently omitted from some questionnaires. ‡This question was left blank in one subject's booklet.

Results and Discussion

Table 6 shows the choices made by the subjects in each condition for each of the feature pairs. These data confirm both the presence of a general trend toward choosing questions with an extreme probability over more symmetrical questions and a preference for extremity given the hypothesis in particular. Of the six significant preferences, all are consistent with that pattern. Clearly, preferences also vary with the context. Condition differences (planetary vs. medical) are significant for Pairs 2, 3, and 5, and marginally significant ($p < .10$) for Pairs 4 and 6. These differences indicate that question choice may be influenced by task-specific characteristics in addition to general information-gathering preferences.

GENERAL DISCUSSION

Much recent work in hypothesis testing has focused on the information-gathering strategies that subjects use, that is, the questions that they ask in order to test one or another focal hypothesis (Baron et al., 1988; Devine et al., 1990; Klayman & Ha, 1987; Skov & Sherman, 1986). The present paper focuses on the subsequent stage in the hypothesis-testing process: how people draw inferences and revise beliefs when they receive the answers to these questions.

We proposed that, although people may be sensitive to differences in the diagnostic values of different questions (Bassok & Trope, 1984; Skov & Sherman, 1986; Trope & Bassok, 1983), they are not very sensitive to differences in the diagnostic values of particular answers to a question. Thus, people may recognize that if you are trying to distinguish Gloms from Fizos, it is better to check a 90-50 feature (i.e., one that is present in 90% of Gloms and 50% of Fizos) than a 70-50 feature. However, they may not recognize that, in both cases, absence of the feature (a "no" answer) is stronger evidence of a Fizo than presence of the feature (a "yes" answer) is of a Glom. If, as we proposed, subjects generally treat different answers to the same question as though they carry similar amounts of information, they will often revise their initial beliefs inappropriately. Their revisions may be either more or less than the normative (Bayesian) prescription, depending on the question asked and the answer received.

Insufficient sensitivity to the diagnostic value of specific answers was demonstrated for two different kinds of judgments. These judgments involved inferences about the category membership of an individual given a single datum and the likely makeup of the underlying population, given data from a random sample. In both cases, the subjects showed little or no sensitivity to the differential diagnosticity of different answers to the same question. The subjects' population estimates (Experiment 2) provide some striking examples of how this can affect inferences drawn from data. For instance, when data concerning a 90-50 feature were consistent with a population of 38 Gloms out of 100 (producing 13 "yes" answers and 7 "no" answers), the subjects estimated 64 Gloms; a 3/8 minority was judged to be a 5/8 majority.

One possible explanation for insufficient sensitivity to answer diagnosticity is that subjects focus on the difference between the probabilities of the feature in each group. This is very useful in selecting questions: The difference between the percentage of members in each group that have a feature is a good approximation of the informativeness of asking about that feature (e.g., with respect to expected value of information or expected log likelihood ratio). Thus, questions about 90-50 features and 50-10 features are about equally diagnostic, and both are more diagnostic than a question about a 70-50 feature. To judge the diagnosticity of an individual answer, however, one must abandon one's focus on the difference in probabilities in favor of one of two likelihood ratios. Thus, a "yes" answer to a 90-50 item is far less diagnostic than either a "no" answer to that question or a "yes" answer to a 50-10 question. Although the difference in probabilities is 40 in each case, the relevant likelihood ratio is 9 to 5 in the first case and 5 to 1 in the last two. Previous research on Bayesian inference has found that people do not generally have an accurate intuition about the appropriate role of likelihood ratios in the revision of beliefs (see Fischhoff & Beyth-Marom, 1983), and that seems to be the case here, as well.

In particular, insensitivity to answer diagnosticity may be related to the heuristic of "representativeness" (Kahneman & Tversky, 1972; Tversky & Kahneman, 1982b). According to representativeness, people judge the likelihood that an individual is a member of a group according to how well the individual's characteristics match those of the average or prototypical group member. In a similar vein, people may judge the impact of a datum by comparing how strongly representative it is of one group versus the other. For instance, if a feature is present in 90% of Gloms and 50% of Fizos, then a "yes" answer is very representative of Gloms and neither representative nor unrepresentative of Fizos. Thus, the answer is considerably more representative of Gloms than of Fizos and constitutes good evidence in favor of the Glom hypothesis. Similar reasoning would prevail if a "no" answer were received. A "no" is very unrepresentative of Gloms and again neutral with regard to Fizos. Because a "no" is just as unrepresentative of Gloms as a "yes" is representative, the two answers may be seen as about equally valuable in distinguishing the two groups.

Given the evidence for insufficient sensitivity to answer diagnosticity, what are the likely overall effects on peoples' judgments about hypotheses? As mentioned earlier, that depends on the questions people ask. Thus, we investigated aspects of subjects' information-gathering processes that might affect the impact of insensitivity to answer diagnosticity. For instance, symmetrical questions (70-30, 20-80) are not prone to the inferential errors we document, because "yes" and "no" answers are equally diagnostic. We found, however, that people generally prefer questions for which one group has an extreme probability of possessing the feature (e.g., .90 or .10) over more symmetrical questions of equal diagnosticity. The preferred asymmetrical questions are also the ones for

which one answer is markedly more diagnostic than the other and are thus the questions most prone to error because of insensitivity to answer diagnosticity.

In light of previous research in hypothesis testing, we were also interested in the possibility that insensitivity to answer diagnosticity might be linked to a bias in favor of the subjects' focal hypothesis. Insensitivity to answer diagnosticity does not by itself imply any such tendency toward hypothesis preservation. When one of two hypotheses is seen as the target for testing, insensitivity could lead to either overconfidence or to underconfidence in that target hypothesis. For example, when subjects learn that an individual has a feature found in 90% of the target group and 50% of the nontarget group, they will overestimate the strength of this relatively weak evidence in favor of their hypothesis. If the feature is found in 50% of the targets and 10% of the nontargets, though, they will underestimate the strength of the relatively strong confirming evidence of a "yes" answer.

However, insensitivity to answer diagnosticity can produce a predominance of errors in favor of the hypothesis if certain kinds of questions are favored over others. In particular, this will happen if subjects tend to choose questions for which the hypothesis-confirming answer is less diagnostic than the disconfirming answer. Then, not appreciating the differential diagnosticity of the two answers, subjects will give too much weight to confirming evidence relative to disconfirming evidence, producing an overall bias toward the hypothesis. Previous research (Devine et al., 1990; Skov & Sherman, 1986) suggests that subjects' choice of questions is consistent with this pattern, and we confirm that finding in Experiment 3. People's preference for extreme-probability features is affected by which is considered to be the focal hypothesis and which the alternative. People tend to ask about features that are extremely common or uncommon given the focal hypothesis, in preference to those with extreme probabilities given the alternative. Thus, for example, they will ask about features present in 90% of the target group and 50% of the nontarget group, in preference to features present in 50% of the targets and 10% of the nontargets. The upshot is that subjects tend to choose questions for which the answer that favors the hypothesis (be it "yes" or "no") is more frequent and less diagnostic than the answer that favors the alternative (see Note 3).

There are several possible explanations for this preference. It has been suggested that people are motivated to avoid questions that are likely to provide disconfirming answers. This might be because people dislike being proven wrong (e.g., Snyder & Swann, 1978; Wason, 1960) or because it is harder to process negative information (see Chase & Clark, 1972, on sentence verification and Wason & Johnson-Laird, 1972, on deductive reasoning). Thus, subjects may prefer questions with extreme probabilities given the hypothesis exactly because they are the questions for which hypothesis-favoring answers are most frequent. However, the best explanation seems to be that people believe these questions to be more

informative. The focus on extreme probabilities given the hypothesis may reflect a kind of representativeness-based judgment similar to that described earlier. If the target group of the hypothesis has a given feature 90% of the time, then having that feature is highly representative of the group and not having it is highly unrepresentative. It may thus seem that features with extreme probabilities given the hypothesis will identify an individual as a member or nonmember with high probability. On the other hand, if a feature is possessed by 50% of the target group, neither having nor lacking the feature is very representative of the target group, so that feature may be seen as uninformative with regard to group membership. Indeed, other things being equal, extreme-probability features are more informative. Unfortunately, other things (in particular, the frequency of the feature given other hypotheses) are not always equal. The tendency to focus on the probability of the datum given the hypothesis, with relative inattention to its likelihood given the alternative, has been referred to as "pseudodiagnosticity" (Doherty, Mynatt, Tweney, & Schiavo, 1979; Fischhoff & Beyth-Marom, 1983).

The findings of our question-selection studies indicate that subjects choose questions in a way that can skew the effects of insensitivity to answer diagnosticity in the direction of the focal hypothesis. The Appendix provides two examples of this process. The illustrations model a hypothetical subject who is attuned to differences in the diagnosticity of questions but who treats "yes" and "no" answers to any given question as being of equal diagnosticity. The hypothetical subject chooses a set of questions that approximates the choice patterns we observed in Experiment 3. The illustrations show that the kind of question-asking preferences we observed can substantially influence the effects of insensitivity in the direction of overestimates of the probability of the focal hypothesis. Note that this hypothesis-preservation effect would not result from the pattern of choosing questions alone, nor from insensitivity to answer diagnosticity in revising beliefs, but only from the combination of the two.

In conclusion, the findings of our studies suggest something about people's fundamental understanding of diagnosticity. People may, to a large extent, see diagnosticity as more a function of the question asked than of the answer received. This differs considerably from the Bayesian view of the world and, as we demonstrate, can lead to systematic errors in belief revision. Our findings also illustrate the importance of looking at a phenomenon such as insensitivity to answer diagnosticity in relationship to the broader context of processes at different stages of hypothesis testing. These processes include gathering information (selecting questions to ask), assessing the information value of data (estimating the confidence associated with individual answers), and drawing inferences based on sets of data (estimating the makeup of underlying populations). People conform to normative statistical models in some aspects of hypothesis testing (e.g., sensitivity to the diagnostic value of questions). In other aspects, peo-

ple show preferences that are not dictated by normative considerations but do not necessarily lead to errors in judgment (e.g., preferences for questions for which the likely answer confirms the hypothesis). And in some aspects, people clearly deviate from normative standards (e.g., insufficient sensitivity to the diagnostic value of different answers to the same question). Moreover, judgments and preferences at early stages in the hypothesis-testing process can combine with judgments at later stages to influence final inferences and beliefs and to produce effects that do not arise from any single component. Our findings suggest, for example, that people may arrive at an unwarranted belief that the world is as it was hypothesized to be—not necessarily because of a motivation to perceive the expected but because of a particular sequence of judgments and inferences, some normatively appropriate, some not.

REFERENCES

- BAR-HILLEL, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, **44**, 211-233.
- BARON, J. (1985). *Rationality and intelligence* (Chapter 4, pp. 130-167). Cambridge: Cambridge University Press.
- BARON, J., BEATTIE, J., & HERSHEY, J. C. (1988). Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty. *Organizational Behavior & Human Decision Processes*, **42**, 88-110.
- BASSOK, M., & TROPE, Y. (1984). People's strategies for testing hypotheses about another's personality: Confirmatory or diagnostic? *Social Cognition*, **2**, 199-216.
- BEACH, L. R. (1968). Probability magnitudes and conservative revision of subjective probabilities. *Journal of Experimental Psychology*, **77**, 57-63.
- BIRNBAUM, M. H., & MELLERS, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality & Social Psychology*, **45**, 792-804.
- BORGIDA, E., & BREKKE, N. (1981). The base-rate fallacy in attribution and prediction. In J. H. Harvey, W. J. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 3, pp. 63-95). Hillsdale, NJ: Erlbaum.
- CHASE, W. G., & CLARK, H. H. (1972). Mental operations in the comparison of sentences and pictures. In L. Gregg (Ed.), *Cognition in learning and memory* (pp. 205-232). New York: Wiley.
- DEVINE, P. G., HIRT, E. R., & GEHRKE, E. M. (1990). Diagnostic and confirmation strategies in trait hypothesis testing. *Journal of Personality & Social Psychology*, **58**, 952-963.
- DOHERTY, M. E., MYNATT, C. R., TWENEY, R. D., & SCHIAVO, M. D. (1979). Pseudodiagnosticity. *Acta Psychologica*, **43**, 111-121.
- EVANS, J. ST. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ: Erlbaum.
- FISCHHOFF, B., & BEYTH-MAROM, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, **90**, 239-260.
- KAHNEMAN, D., & TVERSKY, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, **3**, 430-454.
- KLAYMAN, J., & HA, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis-testing. *Psychological Review*, **94**, 211-228.
- SHERMAN, S. J., & CORTY, E. (1984). Cognitive heuristics. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (Vol. 1, pp. 189-286). Hillsdale, NJ: Erlbaum.
- SKOV, R. B., & SHERMAN, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology*, **22**, 93-121.
- SNYDER, M., & SWANN, W. B., JR. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality & Social Psychology*, **36**, 1202-1212.
- TROPE, Y., & BASSOK, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality & Social Psychology*, **43**, 22-24.
- TROPE, Y., & BASSOK, M. (1983). Information-gathering strategies in hypothesis-testing. *Journal of Experimental Social Psychology*, **19**, 560-576.
- TVERSKY, A., & KAHNEMAN, D. (1982a). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153-160). Cambridge: Cambridge University Press.
- TVERSKY, A., & KAHNEMAN, D. (1982b). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84-98). Cambridge: Cambridge University Press.
- WASON, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, **12**, 129-140.
- WASON, P. C., & JOHNSON-LAIRD, P. N. (1972). *Psychology of reasoning: Structure and content*. London: D. T. Batsford.

NOTES

1. In a test of 2,401 pairs of feature probabilities ranging from .02 to .98 in each group, the correlation between difference in probability and expected log likelihood ratio for the question was .978.

2. An analogy might be drawn here to findings in another area of research in Bayesian inference—people's underuse of base-rate information. There, the strongest form of the hypothesis (that people completely ignore all base-rate data) was rejected, but evidence supports the conclusion that people often underweight such information. This systematic underweighting is sufficient to produce significant discrepancies between intuitive and statistical inferences in many situations (e.g., see Bar-Hillel, 1980).

3. Assume that there are two answers and two hypotheses, such that answer A favors the hypothesis, α , and answer B favors the alternative, β . The more diagnostic answer is always the one that has the less extreme probability given the hypothesis it favors. The diagnosticity of A answers can be taken as the likelihood ratio $LR_A = p(A|\alpha)/p(A|\beta)$; for B answers, the likelihood ratio is $LR_B = p(B|\beta)/p(B|\alpha)$.

$$\begin{aligned} \frac{p(A|\alpha)}{p(A|\beta)} &< \frac{p(B|\beta)}{p(B|\alpha)} \rightarrow p(A|\alpha) \cdot p(B|\alpha) < (p(B|\beta) \cdot p(A|\beta)) \\ &\rightarrow p(A|\alpha) \cdot [1 - p(A|\alpha)] < p(B|\beta) \cdot [1 - p(B|\beta)] \\ &\rightarrow p(A|\alpha)^2 - p(A|\alpha) > p(B|\beta)^2 - p(B|\beta) \\ &\rightarrow [p(A|\alpha) - 1/2]^2 > [p(B|\beta) - 1/2]^2, \end{aligned}$$

which is true if (and only if) $p(A|\alpha)$ is more extreme (further from .5) than is $p(B|\beta)$. If subjects select questions for which $p(A|\alpha)$ is more extreme than $p(B|\beta)$, the confirming answer (A) will be less diagnostic than the disconfirming (B).

4. This occurred on 16% of all items, mostly in response to "no" answers. Concern that such answers might reflect confusion or inattention prompted us to exclude those responses. Doing so may have biased the sample toward more attentive or numerate subjects. However, reanalysis using the full sample revealed qualitatively similar results.

APPENDIX

The two cases shown below illustrate the possible combined net effects of extremity bias in question asking and insensitivity to answer diagnosticity. The two cases use the same selection of seven questions, which roughly follows the findings in Experiment 3 regarding preferences for questions about features with extreme probabilities. In each case, the hypothetical subject asks four questions with extreme probabilities given the hypothesis, α , two questions with extreme probabilities given the alternative, β , and one question with no extreme probabilities.

To represent insensitivity to answer diagnosticity, the subject is assumed to treat "yes" and "no" answers to any given

question as being of equal diagnosticity. That diagnosticity is taken to be equal to the diagnosticity of "yes" and "no" answers to a symmetrical question for which the difference between the α and β probabilities is the same. For example, "yes" and "no" answers to a 90-50 question are treated as though they had inverse likelihood ratios, $p(\text{yes}|\alpha)/p(\text{yes}|\beta) = 7/3$ and $p(\text{no}|\alpha)/p(\text{no}|\beta) = 3/7$, as they would if the probabilities were 70-30. Except for this insensitivity, the subject is assumed to be a proper, unbiased Bayesian.

Case 1. This case illustrates a set of data that would be fairly likely to occur if Hypothesis α were true:

1. $p(\text{yes}|\alpha) = .90, p(\text{yes}|\beta) = .50$, answer = yes, favors α .
2. $p(\text{yes}|\alpha) = .85, p(\text{yes}|\beta) = .55$, answer = no, favors β .
3. $p(\text{yes}|\alpha) = .90, p(\text{yes}|\beta) = .45$, answer = yes, favors α .
4. $p(\text{yes}|\alpha) = .10, p(\text{yes}|\beta) = .45$, answer = no, favors α .
5. $p(\text{yes}|\alpha) = .40, p(\text{yes}|\beta) = .10$, answer = no, favors β .
6. $p(\text{yes}|\alpha) = .50, p(\text{yes}|\beta) = .90$, answer = yes, favors β .
7. $p(\text{yes}|\alpha) = .70, p(\text{yes}|\beta) = .30$, answer = yes, favors α .

Assuming prior odds of 1/1, the odds of α given these data should be the product of the likelihood ratios for each answer, $(90/50) (15/45) (90/45) (90/55) (60/90) (50/90) (70/30) = 1.7/1$,

which equals a probability of .63. The insensitive subject would treat the evidence as equivalent to $(70/30) (35/65) (72.5/27.5) (67.5/32.5) (35/65) (30/70) (70/30) = 3.7/1$, a probability of .79.

Case 2. This case illustrates a set of data that would be fairly likely to occur if Hypothesis α were false:

1. $p(\text{yes}|\alpha) = .90, p(\text{yes}|\beta) = .50$, answer = yes, favors α .
2. $p(\text{yes}|\alpha) = .85, p(\text{yes}|\beta) = .55$, answer = yes, favors α .
3. $p(\text{yes}|\alpha) = .90, p(\text{yes}|\beta) = .45$, answer = no, favors β .
4. $p(\text{yes}|\alpha) = .10, p(\text{yes}|\beta) = .45$, answer = yes, favors β .
5. $p(\text{yes}|\alpha) = .40, p(\text{yes}|\beta) = .10$, answer = no, favors β .
6. $p(\text{yes}|\alpha) = .50, p(\text{yes}|\beta) = .90$, answer = yes, favors β .
7. $p(\text{yes}|\alpha) = .70, p(\text{yes}|\beta) = .30$, answer = yes, favors α .

The odds of α with these data should be $(90/50) (85/55) (10/55) (10/45) (60/90) (50/90) (70/30) = 1/10.3$, a probability of .09. The subject's response would be $(70/30) (65/35) (27.5/72.5) (32.5/67.5) (35/65) (30/70) (70/30) = 1/2.3$, a probability of .30.

(Manuscript received June 5, 1989;
revision accepted for publication February 23, 1992.)

5th European Conference for Research on Learning and Instruction Aix-en-Provence, France August 31 - September 5, 1993

The 5th European Conference for Research on Learning and Instruction, organized by the European Association for Research on Learning and Instruction (E.A.R.L.I.), will be held in Aix-en-Provence, France, August 31 to September 5, 1993.

The conference will include invited addresses, symposia, paper sessions, poster sessions (workshops), and demonstrations. The official language will be English.

The main themes will be: Cultural and social aspects • Learning and teaching in specific subject areas, including foreign languages • Text comprehension and text production • Learning disabilities • Classroom interactions and learning processes • New media and information technologies • Evaluation and assessment • Adult learning and continuing education • Learning in natural settings • Knowledge acquisition, problem solving, and metacognition • Developmental processes • Teaching process and teacher thinking • Industrial and professional training • Social, motivational, and cognitive factors • Parental involvement in learning.

The deadline for proposals is August 31, 1992.

The address for proposals and the contact for further information is: Jean-Paul Roux, 5th EARLIE Conference Secretariat, U.F.R. de Psychologie et Sciences de l'Education, Université de Provence, 29 Avenue Robert Schuman, 13621 Aix-en-Provence Cedex, France (Telephone, +33-42 27 03 08; Fax, +33-42 20 59 05).