ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES 42, 88–110 (1988)

Heuristics and Biases in Diagnostic Reasoning

II. Congruence, Information, and Certainty

JONATHAN BARON, JANE BEATTIE, AND JOHN C. HERSHEY

University of Pennsylvania

Six experiments were carried out to examine possible heuristics and biases in the evaluation of yes-or-no questions for the purpose of hypothesis testing. In some experiments, the prior probability of the hypotheses and the conditional probabilities of the answers given each hypothesis were elicited from the subjects; in other experiments, they were provided. We found the following biases (systematic departures from a normative model), and interviews and justifications suggested that each was the result of a corresponding heuristic: *Congruence bias.* Subjects overvalued questions that have a high probability of a positive result given the most likely hypothesis. This bias was apparently reduced when alternative hypotheses or probabilities of negative results are explicitly stated. *Information bias.* Subjects evaluated questions as worth asking even when there is no answer that can change the hypothesis that will be accepted as a basis for action. *Certainty bias.* Subjects overvalued questions that have the potential to establish, or rule out, one or more hypotheses with 100% probability. These heuristics are explained in terms of the idea that people fail to consider certain arguments against the use of questions that seem initially worth asking, specifically, that a question may not distinguish likely hypotheses or that no answer can change the hypothesis accepted as a basis for action.

Much of the thinking that occurs in daily life and in professions such as science and medicine involves the formulation and testing of hypotheses. Given a set of one or more hypotheses, there are often several questions that might be asked—several pieces of evidence that might be requested—to test these hypotheses. If one assigns prior probabilities to hypotheses and conditional probabilities to each possible answer given each hypothesis, one may calculate the normative value of each question for choosing the correct hypothesis. We do not normally calculate when

Hershey is in the Department of Decision Sciences, The Wharton School. Baron and Beattie are in the Psychology Department. Baron and Hershey are also senior fellows in the Leonard Davis Institute of Health Economics. This work was supported by N.I.M.H. Grant MH37241 (J.B.) and N.S.F. Grant SES82-18565 (J.C.H.). We thank Carol A. Smith for help in conceptualizing the stimuli for Experiments 1–3, and Robert Sternberg and anonymous reviewers for helpful comments. Address correspondence to J. Baron, Psychology Department, University of Pennsylvania, 3815 Walnut St., Philadelphia, PA 19104-6196.

0749-5978/88 $3.00

we evaluate questions, however. Instead, we seem to use a variety of heuristics, some of which may be more effective than others in approximating the result we desire, the discovery of the most useful question. Each heuristic involves attention to some features of a question and inattention to other features.

In considering what heuristics be used, let us begin by postulating a *proper* heuristic (Baron, 1985), which might come reasonably close (without actually calculating) to the normative model defined by probabilities. We consider the case where each question has two possible answers, positive and negative. A good proper heuristic would be to evaluate the proposed question according to its ability to meet three conditions:

1. a positive (or negative) answer is likely given the most likely hypothesis;
2. a positive (or negative) answer is unlikely given other hypotheses; and
3. the hypothesis to be accepted as a basis for action depends on the answer.

The second condition requires consideration of alternative *hypotheses,* and the third condition requires consideration of alternative *answers* and their implications for action. The parenthetical "or negative" implies that conditions 1 and 2 should be applied to negative answers as well as positive ones.

Baron (1985) has proposed that avoidable errors in thinking may result from failure to consider evidence (arguments) against an initial possibility. Such may be the case in hypothesis testing. A person might use the first condition as a way of generating a possible question (possibly for positive answers only). A search for alternative explanations of anticipated answers (as recommended by Platt, 1964; Popper, 1962; and others) would amount to a search for evidence against the initial possibility (question); such a self-critical search would be required in order to satisfy the second condition. Failure to undertake such a search would lead to selection (or overevaluation) of questions whose most likely answer is consistent with a favored hypothesis. We shall call such a heuristic, in which a person attends only to the first condition, the *congruence* heuristic. The use of this heuristic would seem to be especially likely when people have only a single hypothesis in mind. Failure to go beyond the first condition may also result from a general tendency to think too little, to end the search for evidence too early (Baron, 1985; Baron, Badgio, & Gaskins, 1986).

If a question is likely to give different answers under different hypotheses, it still may be worthless, if it cannot possibly change one's best guess about which hypothesis should be accepted as a basis for action—

when it is clear that such a best guess is one's only appropriate goal. For example, if a physician plans to treat a disease even if the result of a (fallible) test for that disease is negative, the test is worthless. Failure to consider the third condition, another failure to seek counterevidence against an initial possibility, can lead to the favoring of worthless questions. We define the *information* heuristic as one in which people attend to the first two conditions but not the third.

Several accounts have suggested that use of an information heuristic might contribute to the performance of medical tests that are expensive, dangerous, and unnecessary (Allman, Steinberg, Keruly, & Dans, 1985; Bursztajn, Feinbloom, Hamm, & Brodsky, 1981, Chap. 1; Elstein, Shulman, & Sprafka, 1978, p. 207), and, in corporations and institutions, to the gathering of data that have little relevance to the decision at hand (Feldman & March, 1981). However, we know of no convincing demonstrations of this heuristic in the laboratory or elsewhere. Note that in order to demonstrate this error is this heuristic, we must show that subjects understand that their ultimate goal is to decide on an action.

In considering the first two conditions, people might adopt a simplifying heuristic, in which they ask simply how likely a question is to establish or rule out one or more hypotheses with 100% certainty. We call this the *certainty* heuristic. At times, this desire for certainty is of paramount importance (e.g., in "pure science; see Baron, 1985, Chap. 4). However, there are other situations in which a best guess is exactly what we want. A question that *might* yield certainty (with a low probability) can have less value than a question that increases the chances of guessing correctly. Confusion about the goal might be one reason subjects would prefer questions that can yield certainty. Another possible reason is that people seem inclined to overweigh certainty in their decisions (Kahneman & Tversky, 1979), and we might therefore anticipate this tendency in people's choice of questions to ask.

Each heuristic leads to corresponding *bias* that will involve over-or under attention to certain features of a question. Congruence bias involves *over*attention to the probability of a "yes" answer assuming that the *favored* hypothesis is true.[1] Information bias involves *under*attention to the probability that actions will differ for different possible answers. Certainty bias involves *over*attention to the probability of ruling some hypothesis in, or out, with certainty.

Evidence consistent with the existence of congruence bias is extensive,

[1] The term *confirmation bias* has been used to mean what we mean by congruence bias (see Fischhoff and Beyth-Marom, 1983; Tweney, Doherty, and Mynatt, 1981), but it has also been used to describe simply choosing a question that is unlikely (or unable) to falsify one's favored hypothesis (e.g., Wason, 1968) or simply being unresponsive to evidence against a favored hypothesis (e.g., Mynatt, Doherty, and Tweney, 1978).

but often there are other interpretations of results, or the generality of the results is doubtful (see Baron, 1985, Chap. 4; Fischhoff & Beyth-Marom, 1983, for reviews). However, one study is particularly relevant here. Shaklee and Fischhoff (1982) gave subjects an event, such as, "Diane rode her bike to work;" some possible causes, such as, "Diane's car wouldn't start;" and, "There were no parking places on campus;" a fact implicating one of the causes, such as, "Even in bad weather, she rode her bike;" and some questions that might be asked, such as, "Had Diane's car been giving her trouble lately?" and "Did Diane ride her bike for convenience?" Subjects were asked which question they "would most like to have answered in trying to explain" the initial event. Subjects strongly tended to choose the question most likely to yield a "yes" answer if the implicated cause were true. This by itself might not be a bias, for the chosen question might still have been the best one. However, in one experiment, one of the causes was labeled as already known (and the implicating fact was omitted). Subjects were told that the alternative causes were not mutually exclusive. Subjects still tended to ask the question corresponding to the known cause. This is such an extreme form of congruence bias that one wonders whether the subjects understood the instructions. Accordingly, in connection with the present experiments, we gather justifications or verbal protocols so that understanding may be checked.

In the present experiments, we seek evidence for the use of the heuristics described above in situations in which they lead to bias, i.e., conflict with a normative model. Comparison across the experiments allows us to ask what conditions affect the biases, and how they might be avoided.

The Normative Model

Before we describe the experiments, we present a normative model (a model of what we should do if we had the time and inclination to calculate) that applies to all our experiments. Consider a situation in which we must evaluate a question (e.g., a diagnostic test in medicine) for the purpose of deciding which of three hypotheses (e.g., diseases) is most likely to be true. Only one question may be asked, and then a decision must be made about which hypothesis to accept (act on). We care only about being correct in this decision, so we can assume that there is utility of 1 for accepting the correct hypothesis and 0 for accepting any other. Our best guess is always the most likely hypothesis, given whatever knowledge we have. Expected utility is therefore equivalent to the probability of the most likely hypothesis. The expected utility of a question is the probability of guessing correctly after asking the question (calculated

before the answer is known) minus the probability of guessing correctly without asking the question (Baron, 1985, Chap. 4; Savage, 1954, Chap. 6).

Table 1 shows the application of this principle to the cases we examine here. H_i (the "hit rate" for hypothesis i) is the probability that the answer is positive, given that hypothesis i is true, and P_i is the prior probability of hypothesis i (before any question is asked). If we do not ask the question, we accept the most likely hypothesis, and our probability of being correct is max (P_i), the largest value in the bottom row of Table 1. If we ask the question, there could be one of two answers, yes or no. We will accept the most likely hypothesis given the answer. We may think about this in terms of the joint probabilities $(P_i)(H_i)$ and $(P_i)(1 - H_i)$. (By definition, prob(positive & i) = (prob(i))(prob(positive/i)) = $(P_i)(H_i)$; likewise for prob(negative & i).) If the answer if yes, we will accept the most likely hypothesis, and the overall probability of this happening is max$[(P_i)(H_i)]$, the largest value in the first row. If the answer is no, the overall probability of being correct is max$[(P_i)(1 - H_i)]$, the largest value in the second row. Thus, the expected utility after asking the question and getting the answer (calculated without knowing what the answer will be) is max$[(P_i)(H_i)]$ + max$[(P_i)(1 - H_i)]$, and the overall utility of the question, relative to simply accepting the most likely hypothesis, is max$[(P_i)(H_i)]$ + max$[(P_i)(1 - H_i)]$ − max P_i. To take an example, if the first row of Table 1 is .4 .0 .2, and the second row is .1 .3 .0, then the third row has to be .5 .3 .2. In this case, we would have a .5 probability of guessing correctly without asking, and a .7 probability (.4 + .3) probability of guessing correctly after the answer, so the expected utility of the question is .2. If the second row were .3 .1 .0, we would have a .7 chance of guessing correctly with or without the answer, and the question would be worthless even though it would help to distinquish between hypotheses 2 and 3 as second choices.

TABLE 1

PROBABILITIES RELEVANT TO THE NORMATIVE MODEL FOR THE VALUE OF A QUESTION ASSUMING THREE HYPOTHESES AND TWO POSSIBLE ANSWERS TO THE QUESTION

Answer	Hypothesis 1	2	3
yes	P_1H_1	P_2H_2	P_3H_3
no	$P_1 (1 - H_1)$	$P_2 (1 - H_2)$	$P_3 (1 - H_3)$
P_i	P_1	P_2	P_3

Note. P_i is the prior probability of hypothesis i. H_i is the conditional probability of a yes answer given hypothesis i. The entries in each cell are the joint probabilities of hypothesis and answer.

There are two types of worthless information. The first type can affect the probability of irrelevant alternatives, those we would not accept in any case. The second type can raise or lower the probability of the favored hypothesis, but cannot lower the probability enough so that we would not accept the hypothesis as a basis for action. These two types occur together when elimination of an unlikely alternative increases the probability of the favored hypothesis. Note that an increase in the probability of the favored hypothesis from one answer does not make us more likely to accept the correct hypothesis after obtaining that answer, for we would accept the same hypothesis with any other answer.

EXPERIMENTS 1–3

The first three experiments concern the congruence bias primarily. Subjects are asked to evaluate questions that they might ask in realistic situations. Because each situation is familiar, subjects may be asked for their own prior and conditional probabilities, after they evaluate the questions. We may then enter these personal probabilities in our normative model, and look for sources of bias.

We adapted the technique used by Skaklee and Fischhoff (1982). We gave subjects a brief description of a situation in which the task was to discover the cause of some event. Unlike the situations used by Shaklee and Fischhoff, ours were designed to encourage the subject to consider mutually exclusive causes. The use of events with mutually exclusive causes made it easier for us to fit the normative model. Most events were rare and undesired, the sort of event that would probably have only a single cause. The description was designed to make one particular cause seem more plausible than others. In Experiment 1, we asked subjects to write down the hypotheses that occurred to them and to evaluate three questions that could help discover the cause. After doing this for 20 descriptions, they assigned probabilities to their hypotheses and to the occurrence of yes or no answers to each question given each hypotheses. From these probability assignments, we could calculate the normative value of each question, given the subject's own beliefs, and we could look for sources of deviation from the normative model.

In Experiment 2, we provided a single hypothesis about the cause of the event, and the subject was asked to evaluate the question with respect to that hypothesis. This was to ask whether the biases found in Experiment 1 were the result of the subjects' attachment to hypotheses they thought of themselves.

In Experiment 3, we provided two hypotheses instead of one, in order to ask whether this would reduce the biases found in Experiment 2. Congruence bias might be caused by subjects' failure to ask themselves

whether a result could also be produced by a hypothesis other than the one under consideration.

METHOD

Subjects were solicited by a sign placed on the main walkway of the University of Pennsylvania; most were undergraduate students, some were graduate students. They were paid $3.75 per hour.

Subjects filled out a written questionnaire. For Experiment 1, each item described an event and asked a question, for example, "Karen was 6 years old. Her parents noticed that she was covered with little red spots. She had a mild fever. WHAT MIGHT BE WRONG WITH KAREN?" The subjects then wrote down any hypotheses that occurred to them about the cause of the target event. (The instructions explicitly mentioned hypothes*es*, although most subjects mentioned only one.) The subjects turned the page and then followed these instructions (given at the beginning): "Next, you will find three questions, each of which might give you information about the cause of the target event. We would like you to rate the value of each of the questions on a scale from 0 to 100. If you think that asking that particular question has no value at all, then you should give it 0. If the question would give you all the information you needed to ascertain the true cause, then it should receive 100. When you rate the value of the question, assume that it is the *only* question you could ask." For example, the questions for the event above are: "1. Did one of Karen's friends have measles? 2. Was Karen allergic to mosquito bites? 3. Had Karen been sitting out in the sun all day? All questions could be answered yes or no. Subjects were told to imagine that the facts given to them have occurred to them spontaneously, and, if they asked, they were told to imagine that the people involved were people like themselves.

After answering these questions for 20 different events (hence 60 questions), subjects were asked to review the entire questionnaire and assign probabilities to each of their hypotheses. In addition, they were to indicate the probability of a yes answer to each question, assuming each of their hypotheses to be true. For this purpose, "all other possibilities" was considered as another hypothesis. Thus, a subject who had one hypothesis would have to provide six conditional probabilities for each event: the probabilities of yes answers to each question assuming the hypothesis to be true, and the probabilities assuming the hypothesis to be false.

Fourteen subjects were tested. The entire procedure took up to 3 hr, especially for the five subjects who gave more than one hypothesis, and subjects occasionally were asked to make the ratings in a second session. Because three of the multiple-hypothesis subjects did not return for the

second session, we analyze only the single-hypothesis problems from the nine subjects who gave only a single hypothesis on all or almost all problems.

In Experiment 2, a hypothesis was provided along with the description of the event. For example, "This winter Jennifer returned from her vacation in Switzerland with a broken leg. You suspect that she broke her leg skiing." (The questions for this item were: "1. Had she been mountain climbing? 2. Had she been mugged? 3. Had she been skiing?") There were 15 items (3 questions each, for a total of 45 questions) rather than 20. Instructions were as before, except that subjects were asked to evaluate the question in terms of whether it would "improve your chance of guessing correctly whether your idea [hypothesis] is true or not." Subjects provided probabilities for the hypothesis and conditional probabilities for yes answers given the hypothesis and its converse.

Data from 10 subjects were used. One additional subject gave conditionals that always added to one for each question, an indication of misunderstanding.

Experiment 3 was identical except that two hypotheses were given, the number of items was cut to 10 (30 questions), and subjects were asked to evaluate the questions with respect to whether each question would help them guess which hypothesis, if any, was true. An example of an item is: "Everyone was enjoying the fraternity party when suddenly the fire alarm went off. Hypothesis 1: It was a prank. Hypothesis 2: A lighted cigarette butt had started a fire. Questions: (a) Was it April 1? (b) Was someone smoking nearby? (c) Was it a hot, dry evening?" Subjects gave probabilities for each hypothesis and conditional probabilities for each question assuming that each hypothesis was true and that neither was true.

Data from 12 subjects were used; no subject was omitted.

RESULTS AND DISCUSSION

To look for congruence bias, we first fit a quasi-normative model to each question, for each subject (in all three experiments). By quasi-normative, we mean a model like the normative model except that it allows certain departures other than those or primary interest. We then predicted the subject's ratings from the model using linear regression, and we looked for correlated of the residual. For example, we asked whether deviations could be accounted for by assuming that a subject favored questions with high conditional probabilities of yes answers given the most likely hypothesis. If so, the residual would correlate with the hit rate (H, the probability of a yes answer assuming the hypothesis to be true), and we could conclude that the subjects' answers were consistent with a congruence bias.

TABLE 2
ILLUSTRATION OF THE FITTING OF MODELS FOR EXPERIMENTS 1–3

Joint probabilities [e.g., p(true & yes) = $P \cdot H$ = (.85) (.10) = .085]

Question	Answer	Assuming P = .85		Assuming P = .50	
		True	False	True	False
1	yes	.085	.12	.05	.40
	no	.765	.03	.45	.10
3	yes	.595	.06	.35	.20
	no	.255	.09	.15	.30

Model predictions	Question 1	Question 3
Normative model: max[$P \cdot H$, (1 − P)F] + max[P(1 − H), (1 − P)(1 − F)] − max[P,1 − P]	.035	.00
Equal priors: same as normative but with P = .50	.35	.15

Note. The example is from a subject's response in Experiment 2. "You open the refrigerator one night and discover that there is water on the floor. You think that the regrigerator has stopped working and the ice has melted." P = .85. Question 1: "Had there been a large jug of water in the refrigerator?" Rating = 10, H = .10, F = .80. Question 3: "Was the refrigerator old?" Rating = 25, H = .70, F = .40.

Table 2 shows how both the normative and quasi-normative models may be fit on the basis of two-by-two tables of joint probabilities: true vs. false for the hypothesis, and yes vs. no for the answer to the question. Here, H represents the probability of a yes answer if the most likely hypothesis is true, F represents the probability of a yes answer if this hypothesis is false, and P represents the probability of the most likely hypothesis; all these numbers are provided by the subject.[2,3] The normative utility of the question is the maximum joint probability if the answer is yes, plus the maximum if the answer is no, minus the maximum probability of guessing correctly without asking the question (that is, the maximum of P and 1 − P). When the maxima are all in the same column, the normative utility is zero.

The quasi-normative model was equivalent to the normative model except that we assumed that all priors were equal. This model fit the data better than the normative model for all 9 subjects in Experiment 1, 7 of 10 subjects in Experiment 2, and 11 of 12 in Experiment 3.[4] After calculating

[2] P is constant for the three questions following each item.

[3] The most likely hypothesis could be, but rarely was, the complement of the given hypothesis, or the complement of both hypotheses in Experiment 3.

[4] The mean correlations are .25, .33, and .36 for the normative model, and .43, .43, and

TABLE 3
MEAN CORRELATIONS OF MODELS FOR EXPERIMENTS 1–3

Model	Exp. 1	Exp. 2	Exp. 3
1. Equal *P*	.43(8.20)***	.43(5.89)***	.50(13.4)***
2. Normative (1)	.00(−.06)	.04(1.11)	.09(2.05)*
3. *Mx* (1,2)	−.04(−.73)	−.04(−.97)	.02(0.30)
4. *H* (1,2,3)	.30(7.23)***	.33(6.74)***	.19(4.27)***
5. *F* (1,2,3)	.03(0.59)	−.04(−.75)	.08(1.40)

Note. All correlations except those of the normative model are correlations with residuals. Numbers in parentheses after each model name indicate the other models whose effects have been removed by residualizing. Numbers in parentheses after each mean are *t* values across subjects. Significance levels are one-tailed: *, .05; **, .01; ***, .001.

the quasi-normative utility of each question, we regressed (linearly) each subject's ratings on the predictions of the model for that subject. We use residuals as our measure of error, rather than absolute deviations; we thus allow subjects to depart from the model by any linear transformation.[5]

We then correlated these residuals with the normative model, to determine whether priors were taken into account appropriately at all, and we residualized once again, so that any contribution of the normative model was removed. Next, we correlated the last residual with *Mx*, the probability of the most likely hypothesis, with the idea that subjects might be ignoring or underweighing the probability of guessing correctly without asking the question. Once again, we removed any such effect by computing the residual. We correlated this final residual with *H* and *F*.[6] A congruence bias would exist if the *H* correlation is positive and the *F* correlation is near zero.

The results of all three experiments are shown in Table 3. For Experiments 1 and 2, the normative model does not contribute significantly (across subjects) once the quasi-normative model has been fit. Only for

.50 for the quasi-normative model, for the three experiments, respectively. Note that, according to this model, it is impossible for a question to have 0 value so long as hit rates differ for the different hypotheses; thus, this model is consistent with the apparent information bias to be described.

[5] For four questions in Experiment 2 and three in Experiment 3, the question was phrased in such a way that the confirming answer would actually be "no." For example, when the hypothesis at issue was the light went out because the power failed, the question that might show congruence bias was, "Were the lights on in the house across the street?" Clearly, one might well ask this question to confirm the hypothesis in question. Thus, in the analysis, the answers to these questions were reversed. The results were substantively identical when these changes were not made.

[6] To compute *F* in Experiment 3, we took priors of the two other hypotheses into account in computing the overall probability of a yes answer if the most likely hypothesis was false.

Experiment 3 does the normative model make a contribution (however, its contribution is not significantly greater than that in Experiment 2). *Mx* does not correlate with the residual at all in any of the experiments.

In all three experiments the final residual correlates significantly with *H*. The example in Table 2 is typical, in that both the normative and quasi-normative models stipulate that the first question is more valuable than the third (e.g., .35 for the first, .15 for the third, for the quasi-normative model). Yet, the subject gave a higher mean rating to the third (25 vs 10), presumably because it has a higher hit rate (.70 vs .10) for the most likely hypothesis.

These correlations alone could occur if subjects simply gave high ratings to questions likely to yield positive answers. Such a tendency would produce positive correlations for *F* as well, and these are generally not found. In fact, the *H* correlation is higher than the *F* correlation in Experiment 1 ($t(8) = 2.97$, $p < .05$, one-tailed), Experiment 2 ($t(9) = 4.22$, $p < .005$), and Experiment 3 ($t(11) = 2.16$, $p < .05$). These differences would not be found if the *H* correlation were the result of a tendency to consider only the probability of a positive answer.

The results could also occur if subjects overvalued high hit rates *and* low false-alarm rates. In this case, the *F* correlations would be significantly *negative,* which they were not. In fact, the *H* correlations were higher than the *F* correlations with their sign reversed (Experiment 1, $t(8) = 6.72$, $p < .001$; Experiment 2, $t(9) = 5.75$, $p < .001$; Experiment 3, $t(11) = 2.98$, $p < .01$).[7,8]

It might still be argued that the ratings are affected by *both* the overall probability of a positive answer and the extremity of *H* and *F*. If this is the case, provision of two hypotheses rather than one will have no effect on the bias found. However, the provision of two hypotheses in Experiment 3 reduced the congruence bias that was present in Experiment 2. Experiment 3, compared to Experiment 2 (using only the 10 items common to

[7] Although standard deviations of *H* were higher than those of *F* in Experiments 1 and 2 (means of .098 and .101 for *H* in the two experiments, respectively, 0.75 and .066 for *F*), this difference is unlikely to explain the results. The size of the difference in standard deviation (for each subject) correlated −.24 and .04 with the size of the difference in the (absolute) coefficients for *H* and *F*. Correlations between *H* and *F* were −.05 and −.15.

[8] Congruence bias—overattention to hit rate—might be more extreme when the probability of the favored hypothesis is higher. If so, the correlation of $P \cdot H$ with the residual would be higher than that of $(1 - P)H$. This occurred only in Experiment 2 ($t = 5.52$, $p < .001$; $t = 1.32$ and 1.20 for Experiments 1 and 3, respectively). In all three experiments, both $P \cdot H$ and $(1 - P)H$ correlated significantly with the residual whenever *H* did (and $1 - P)F$, like *F*, never correlates significantly with the residual). In general, then, subjects seem to overvalue questions with a high hit rate given the favored hypothesis and to give little consideration to how strongly that hypothesis is favored.

both experiments),[9] shows a smaller contribution of H ($t(20) = 2.69$, $p < .01$, one-tailed). In addition, the difference between the contributions of H and F is smaller in Experiment 3 ($t(20) = 3.12$, $p < .005$). (Experiments 2 and 3 did not differ in the contribution of F or of the difference between H and the negative of F.) These results are consistent with the effect of the extra hypothesis in Experiment 3 on congruence bias.

The difference between Experiments 2 and 3 suggests that following explanation of the congruence bias: If there is some question whose positive answer is congruent with the best hypothesis, that question will be considered a good one to ask, unless its consistency with other hypotheses is considered. Consideration of alternative hypotheses may lead people to consider the second condition, whether a positive answer is consistent with other hypotheses as well.

Additional support for the existence of a congruence bias is provided by an analysis based on purely ordinal data. Such an analysis is relevant because subjects may translate their ratings and subjective probabilities into numbers nonlinearly, and it is conceivable that such nonlinearities could produce the basic result in Experiments 1 and 2 (although we cannot imagine how). Within each group of three questions in Experiments 1 and 2, we looked for cases in which a subject's ordering of the value of two questions was incorrect given any monotonic increasing transform of the probabilities. Specifically, if i and j index the two questions, respectively, and V represents the value assigned to the question, we looked for cases in which:

A. $1 - H_i > H_j > F_j > 1 - F_i$ and $V_j > V_i$, or

B. $1 - F_i > F_j > H_j > 1 - H_i$ and $V_j > V_i$.

In Case A, question i is normatively at least as valuable as question j, yet j is assigned a higher rating (as in Table 2). If the hypothesis is true, the probability of the answer most congruent with the hypothesis (a negative answer to question i or a positive answer to question j) is higher for question i than for question j, since $1 - H_i > H_j$, and, if the hypothesis is false, the probability of the answer most congruent with the hypothesis is lower for question i than for question j, since $F_j > 1 - F_j$. In Case B, "positive" and "negative" are reversed. (The normative ordering of questions i and j holds regardless of the priors, which are held constant across the three questions in each set.) In Case A, the subject departs from correct ordering of the questions in favor of the question with the high hit rate, and, in Case B, in favor of the question with the high false-alarm rate. The type A violation, but not the type B, is consistent with use of the congruence heuristic.

[9] Comparison with Experiment 1 is inappropriate, as the method was quite different.

In Experiments 1 and 2, nine subjects (six from Experiment 2, three from Experiment 1) showed more type A violations than type B, and only one (from Experiment 1) showed more type B than A ($p < .02$, binomial test). (All subjects had one violation each, except for one who had two of type A and one of type B. In each experiment, all violations were from different items.) We may conclude that the congruence bias shown in these experiments is not the result of a distortion in the representation of subjective probability.

Experiments 1–3 demonstrate information bias as well as congruence bias. According to the normative model, the expected utility of the question was zero for a mean of 36 out of 60 questions in Experiment 1, 24 out of 45 in Experiment 2, and 14 out of 20 in Experiment 3. The mean number of zero ratings in the three experiments was 5, 6, and 5, respectively. In these experiments, the apparent source of information-bias is a tendency to ignore priors (as shown by the superior fit of the equal-priors model to the normative model); when priors are sufficiently high, a question may be unable to change one's best guess.

So far, we have provided no direct evidence for the congruence and information heuristic. Such evidence can be provided by asking subjects to justify their answers in writing or to think aloud in face-to-face interviews. In addition, these data provide a check on whether subjects accept the task as given, and whether they use some other heuristic we have not imagined.

We report direct evidence of the information heuristic in connection with Experiment 4. To look for such evidence of the congruence heuristic, 22 undergraduate subjects were given the following problem and asked for written justifications, and 10 others (premedical students) were interviewed and asked to think aloud:

> A patient has a .8 probability of having Chamber-of-Commerce disease and a .2 probability of Elk's disease. (He surely has one or the other.) A tetherscopic examination yields a positive result in 90% of patients with Chamber-of-Commerce disease and in 20% of patients without it (including those with some other disease). An intraocular smear yields a positive result in 90% of patients with Elk's disease and in 10% of patients without it. If you could do only one of these tests, which would it be? Why?

Normatively, the intraocular smear is the better test. (This is clear immediately if we switch the names "positive" and "negative" for that test; this item involves a kind of "framing effect.") However, a subject who used a congruence heuristic would choose the tetherscopic test, as this test is more likely to give a positive result for the more likely disease. Twelve of the 22 undergraduates chose the tetherscopic test.

Written justifications yielded clear evidence of the confirmation heu-

ristic in subjects who clearly understood the task, e.g., "Tetherscopic exam, because the probability of having C of C is higher—I'd want to know if I have it." The interviews also yielded such evidence, for example, ". . . it makes sense to just do the test for the Chamber of Commerce disease, because he has a better chance of having that instead of the Elk's disease. . . ."

EXPERIMENTS 4–6

In the remaining experiments, we provide probabilities rather than elicit them. To accomplish this, we move to unfamiliar situations, specifically, abstract problems in medical diagnosis. Experiment 4 was set up to look for certainty and (especially) information biases; most tests were normatively worthless. In Experiments 5 and 6, most tests were worthwhile, and several models of tests evaluation could be examined. To determine whether difficulty in combining priors with hit rates was crucial in the biases that were found, Experiment 5 presented joint probabilities of disease states and test results, and Experiment 6 presented the priors and conditional probabilities of test results given diseases separately.

Experiment 4: Method

The subject was asked to evaluate medical tests for deciding which of three diseases (A, B, or C) to treat. The three diseases differed in the probability of producing a chemical called tutone, which is detectable in the blood. They were told, "Whether you do the test or not, you will always treat the most likely disease. If you are wrong, the patient will not recover as quickly, but you will not know this early enough to try another treatment." For each test, the subject was given a table indicating the prior probability of each disease (based on "your examination of the patient and everything you know," the probability that tutone was *present* in patients who had the disease, and the probability that tutone was *absent* in patients who had the disease.[10] The prior probabilities of the diseases were constant across all cases: .64 for A, .24 for B, and .12 for C. The probabilities were explained. The subject was instructed to "rate the value of the test on a scale from 0 to 100, where 0 means that the test is worthless and should not be done and 100 means that the test would remove all doubt about which disease the patient has. A rating above 0

[10] We found it necessary to provide probabilities of negative as well as positive results. In an unreported experiment in which probabilities of positive results only are provided, many subjects evaluated tests on the basis of the overall probability of a positive result, as if such a result were good in itself. This phenomenon was not found in Experiments 4–6.

means that the test has some value in improving the accuracy of diagnosis."

The cases used in this experiment are shown in Table 4. Only in Cases 2 and 9 was the test normatively worth doing. Cases 1 and 10 do not meet conditions 1 and 2, so we would expect 0 ratings on these even from subjects who ignore condition 3. Other comparisons will be discussed in connection with the results.

The cases were presented in numerical order to seven subjects and in reverse order to seven. Two of the latter were omitted from the analysis, both for giving ratings separately for each of the three diseases.

Experiment 4: Results and Discussion

The mean ratings are shown in Table 4. Subjects were generally quite likely to realize that the test was worthless in Cases 1 and 10; only four subjects gave positive ratings for Case 1 and only two for Case 10 (both among those who rated Case 1 positive). Subjects were also sensitive to the normative value of the test: Cases 2 and 9 received the highest ratings, and Case 2 had higher ratings than 4 ($t(11) = 2.50, p < 0.25$; all significant statistical tests for this experiment remained significant when Wilcoxen tests were used instead of t tests and when subjects who gave nonzero ratings to Cases 1 and 10 were eliminated).

One kind of information bias involves information about irrelevant alternatives. Cases 4 and 6 provide information about diseases that will not be treated regardless of the result. The mean rating of these cases was higher than the rating of Case 10, which is matched to Cases 4 and 6 in mean probability of a positive result but which does not meet conditions 1 and 2 ($t(11) = 3.30, p < .005$).

TABLE 4
CASES AND MEAN RATINGS FOR EXPERIMENT 4

Case	H_A	H_B	H_C	Normative value	Mean rating (*SD*)
1	.75	.75	.75	0	21 (32)
2	.00	1.00	.00	24	61 (27)
3	.75	1.00	1.00	0	40 (30)
4	.50	1.00	.00	0	34 (25)
5	.50	.00	.00	0	26 (22)
6	.50	.00	1.00	0	26 (22)
7	.50	1.00	1.00	0	48 (28)
8	.25	.00	.00	0	25 (18)
9	1.00	.00	.00	24	75 (13)
10	.50	.50	.50	0	9 (20)

Note. In all cases, $P_A = .64$, $P_B = .24$, and $P_C = .12$. H_A is the probability of a positive test result given disease A, etc.

A second kind of information bias concerns evidence that can change the probability of a hypothesis that will be accepted in any case. Cases 5 and 7 are more capable than Cases 3 and 8 of changing the probability of such a hypothesis (disease A). The mean rating of Cases 5 and 7 was higher than that of Cases 3 and 8 ($t(11) = 3.86, p < .005$), indicating such a bias. The mean of Cases 3 and 8 was also higher than that of Case 10 ($t(11) = 4.35, p < .001$), presumably for the same reason.

Although subjects show an information bias, four subjects gave (a total of 10) 0 ratings to some of Cases 3–8, often with justifications, such as, "You have to treat the patient as with disease A in either case, with or without tutone present."

Cases 5 and 7 allow the possibility of confirming disease A (the most likely) with *certainty,* while 4 and 6 do not. (If the test is positive in Case 5 or negative in Case 7, the patient is sure to have disease A.) The mean rating for Cases 5 and 7 was significantly greater than the mean for Cases 4 and 6 ($t(11) = 2.67, p < .02$). Thus, there seems to be a certainty bias, at least for the most likely disease.

There is little evidence for congruence bias in the ratings (or in the justifications). Case 9 received higher ratings than Case 2, but the difference is not significant. The absence of this bias is most easily attributed to the provision of multiple hypotheses (as in Experiment 3).

The main finding of this experiment is the presence of information bias. The reality of such a bias is supported by interview data from two additional problems:

> 1. A patient's presenting symptoms and history suggest a diagnosis of globoma, with about .8 probability. If it isn't globoma, it's either popitis or flapemia. Each disease has its own treatment, which is ineffective against the other two diseases. A test called the ET scan would certainly yield a positive result if the patient had popitis, and a negative result if she has flapemia. If the patient has globoma, a positive and negative result are equally likely. If the ET scan were the only test you could do, should you do it? Why or why not?
>
> 2. A [different] patient has a .8 probability of umphitis. A positive Z-ray result would confirm the diagnosis, but a negative result would be inconclusive; if the result is negative, the probability would drop to .6. The treatment for umphitis is unpleasant, and you feel it is just as bad to give the treatment to a patient without the disease as to let a patient with the disease go untreated. If the Z-ray were the only test you could do, should you do it? Why or why not?

In these problems, the test cannot change the physician's course of action and is thus normatively worthless. In Problem 2, the test can increase the physician's confidence in the most likely diagnosis. In Problem 1, the test provides information about unlikely alternatives which will not be treated in any case (and it can increase the probability of the favored hypothesis as well).

Although a majority of the 33 subjects given these problems answered

them correctly, interviews with several subjects who were incorrect gave further evidence for inappropriate use of the information heuristic. Sometimes this heuristic seemed to prevent the subject from understanding the relevance of certain information. In other cases, the subjects, with varying amounts of help from the interviewer, were able to understand their error. The following excerpts—which are representative except for their brevity—also demonstrate that the biases are not the result of misunderstanding or of adding assumptions (e.g., the possibility of performing other tests) that would render our normative model inapplicable.

> Subject 1, problem #1. In this case, the interviewer (Baron) has asked why the test would be worth doing even if it cost $5 (this probe being used to negate the assumption that the test is free of both cost and risk). "Because at least to me, the added information is worth the money [section omitted]. [Interviewer: How is it gonna help treat it?] It will give added information. If you do have globoma, the test means nothing. So giving the test or not giving the test, there's no difference. If you do have one of the other diseases, however, the test will mean something. [What does it mean?] If the test is negative, given that you don't have globoma, you'll have flapemia. If the test is positive, given that you don't have globoma, you'll have P. [So what are you gonna do then?] Well, it helps. It can't hurt. So then . . . I mean . . . You obviously look into it to see if there's globoma, and you try to take other tests whatever. [No, that's the only test you can do.] That's the only test you can do. Then you're in a difficult situation."

This subject does not see the normative answer even when it is made clear to him that he has overlooked the fact that only one test is available. In general, he seeks information for its own sake, without asking why it might be useful.

> Subject 2, #1 [after some initial discussion]. "I'd go ahead and do the ET scan if it were not expensive. [What if it cost $5?] Then I would do it. [Okay. Why?] I understand that . . . Okay, a person coming in has a 20% chance of having popitis or flapemia, right? [Uh, huh.] Now, if I couldn't do any other test, I would like to rule out as much of that 20% as I could. If I ruled out some of that, then the 80% would probably increase, right? So if I could rule out popitis . . . [Why would you want the 80% to increase?] I guess to increase the probability of the patient having that disease. [Okay.] If I understand this correctly, popitis and flapemia both present similarly. [All three present similarly, and the reason you think it's 80% is that globoma is just a lot more common.] So I could do the ET scan and rule out popitis, which, as you say, presents very similarly to globoma. I would have narrowed down my choices to globoma or flapemia."

This subject shows evidence of both information-seeking heuristics: he wants to increase the probability of the most likely disease, and he wants to rule out one of the irrelevant ones.

> Subject 3, #2 [after discussion]. "Okay then my first choice would be to definitely do the Z-ray test. [How come?] Because of its minimal risk or cost to the patient, and . . . its answer will help me decide whether or not to treat the patient for this disease. [Okay, how will it help you decide?] Well if the result is positive, then it's

obvious, then the patient has the disease and I treat them for it. However, if it's negative, then the probability is still greater that the person does have the disease, and I think that I would have to . . . I would have to go with the probability. [So what would you do.] I would treat for the disease. [Okay. So why would you give the test, or would you?] Why would I give the *test?* [Um hm.] I would give the test simply to help me decide, because the test can tell me one choice, one way if not the other. [How will it help you . . .] I see, I see. If I give this test . . . Whether or not I give this test, according to what I just said, either way I would be treating for this disease. So, it . . . yeah . . . that makes the test kind of ineffectual, it makes it irrelevant."

The fact that this subject understood the worthlessness of the test after questioning indicates that his original decision to do the test was not based on misunderstanding or addition of extraneous assumptions.

Subject 4, #2, "Yes, I think you definitely should do it, because . . . it's gonna give you a conclusive result in 80%, and . . . yeh . . . you definitely should do it. [Okay, what does it mean to you to say that the two different kinds of mistakes are equally bad? Is that relevant?] I don't think so. [Okay, what would you do if the test is negative, and the probability is now .6?] The odds are still in your favor to give the treatment. [Now does that affect whether you should give the test?] Okay. Yeh. [So you wouldn't give it.] No. [Was there something you misunderstood that made you not see that the first time? or something you didn't look at?] Yeh, I didn't put the two together. No matter what, you're gonna do the treatment, and I took them apart, and I said, you do this test, then you definitely know, and that's . . . definitely do it. But you're probably gonna do it anyway. The reason I said definitely do it is because it given you a concrete positive result, but I didn't realize you're gonna probably do it anyway . . ."

A number of justifications in the present experiment were like the following (although rarely as thorough): "(Case 4) If tutone is present, I know definitely that the patient does not have disease C. If it is absent, I know for certain he does not have disease B. In that case, I would easily choose A as more significant. If tutone was present, I would have a difficult time deciding whether the patient had disease A or B." The value of certainty may be that it makes the decision easy. When there is no certainty, the choice of diseases must be made on the basis of posteriors or joint probabilities, which may be hard to estimate. In the next experiment, we try to eliminate the estimation problem by providing joint probabilities of result and disease.

Experiment 5: Method

The instructions were essentially the same as those for Experiment 4, except that joint probabilities were provided instead of conditional probabilities of tutone given disease. Specifically, subjects were told, "The second and third columns of numbers show the probability of different combinations of disease and tutone. These six numbers add to 1. For example, out of every 100 cases like the one shown in the table, there will

be 6 who have disease A and tutone, 54 who have disease A and no tutone, and so on." This change ought to make it easier for subjects to decide which treatment they would provide given each test result. The correct treatment is for the disease corresponding to the highest joint probability in each column. If the certainty bias found in Experiment 4 were the result of not being able to make such comparisons when only conditional probabilities (of test result given diseases) are given, that effect would disappear here.

The cases used in this experiment are shown in Table 5. The cases differed in the probability that test results could rule out a disease or establish one (by ruling out two of the three possible diseases) with certainty.

Twelve subjects were run with the cases in order, 8 with the cases reversed. Two of the former and one of the latter were excluded: one for giving all 0 ratings, one for answering separately for each disease, and one for requiring an unusual amount of help from the experimenter.

Experiment 5: Results

We fit two models to each subject's ratings. The numbers corresponding to each model are shown in Table 5. By the *normative* model, all tests had values of either 0, .12, or .24. The other model, the *certainty* model, corresponds to the use of the certainty heuristic. It is the expected number of diseases that will be ruled out by doing the test. For example, in Case 2, a positive result, which occurs with probability .88 (.64 + .24),

TABLE 5
CASES, MODEL PREDICTIONS, AND MEAN RATINGS FOR EXPERIMENTS 5

Case	Conditionals			Model predictions		Mean ratings (*SD*)	
	H_A	H_B	H_C	Normative	Certainty	Exp. 5	Exp. 6
1	.50	.50	.50	.00	—	—	—
2	1.00	1.00	.00	.12	1.12	42 (25)	64 (26)
3	.81	.00	.00	.12	.52	56 (18)	62 (22)
4	.00	1.00	.00	.24	1.24	64 (26)	75 (21)
5	1.00	.50	.00	.12	1.00	41 (28)	52 (20)
6	.00	1.00	1.00	.24	1.64	69 (19)	75 (16)
7	.00	.50	.00	.12	.24	44 (27)	41 (26)
8	1.00	.00	1.00	.24	1.24	65 (25)	69 (18)
9	.00	.00	1.00	.12	1.12	42 (31)	56 (24)
10	1.00	.00	.00	.24	1.64	64 (20)	79 (13)
11	1.00	1.00	1.00	.00	—	—	—

Note. In all cases, $P_A = .64$, $P_B = .24$, and $P_C = .12$. We present conditional probabilities here, although joint probabilities were used in Experiment 5.

rules out disease C, and a negative result, which occurs with probability .12, rules out both diseases A and B, so the expected number of diseases ruled out is $.88 \cdot 1 + .12 \cdot 2$, or 1.12.

Most subjects answered 0 to Cases 1 and 11 and gave much higher ratings to all other cases. To avoid scaling problems, we fit the models to Cases 2–10 only. (The results are qualitatively identical when all cases are used.) We also fit the models to the mean ratings across subjects; these mean ratings are shown in Table 5.

The normative model fit well for most subjects. Its mean correlation with each subject's ratings was .57 ($SD = 0.14$, $t(16) = 16.07$, $p < .001$). The certainty model also correlated with subject's ratings (mean $r = .35$), but the two models correlated .73. The mean partial correlation between the certainty model and the ratings (partialing the normative model for each subject) was $-.10$ (not significantly different from 0). It therefore seems that subjects did not use a certainty heuristic when joint probabilities were available.

Comparison of Cases 3 and 7, and 10 and 4, revealed no evidence for congruence bias. (Information bias cannot be tested here.) Again, the absence of congruence bias may be attributed to the provision of multiple hypotheses.

Experiment 6: Method

This experiment was identical to Experiment 5 except that conditional probabilities (of tutone given disease and of tutone given no disease) were given rather than joint probabilities. It was done to ask whether the results of Experiment 5 would hold in what we take to be the more usual situation.[11] Seven subjects were run in the forward order, eight, backward. One subject in each order failed to provide justifications and was omitted. In addition, one of the forward subjects consistently evaluated tests on the basis of the overall probability of a positive result, and the normative model for this subject correlated $-.01$ with the subject's ratings; this subject was also omitted.

Experiment 6: Results

Table 5 shows the mean ratings. The normative model again correlated well with subject's ratings (mean $r = .56$, $SD = 0.27$, $t(11) = 6.84$, $p < .001$. The certainty model correlated as well as the normative model (mean $r = .56$, $SD = 0.28$). The mean partial correlation between the certainty model and the ratings (partialing the normative model) was .35

[11] In most real situations, the hit rate given a hypothesis is relatively independent of factors that affect the prior probability of the hypothesis, but the joint probability is not independent of these factors. Hence, the hit rate may be easier to estimate.

(SD = 0.35), which was significantly greater than 0 (t (11) = 3.35, $p <$ .001). The partial correlations were also significantly higher in Experiment 6 than Experiment 5 (t (27) = 3.62, $p < .001$). It appears that the certainty heuristic is used primarily when subjects cannot compare joint probabilities directly.

DISCUSSION

We have found evidence for the existence of a number of interrelated biases (departures from the normative model) and heuristics (reasons for favoring a question), and we have obtained some initial evidence concerning the aspects of the situation that call forth or prevent them. Table 6 summarizes the results.

In *congruence* bias, a question is overvalued when it is likely to give a yes answer if the most likely hypothesis is true. This bias is substantially reduced in Experiment 3, where alternative hypotheses are provided, and is absent in Experiments 4–6, possibly also because of the multiple hypotheses. Subjects' justifications indicate that the cause of this bias is a heuristic in which subjects favor questions that meet condition 1 (consistency of a positive result with the hypothesis), without checking conditions 2 or 3, and without considering the significance of negative answers.[12]

One factor that does not appear to play a role is the subjects' attachment to hypotheses they have thought of themselves. If anything, congruence biases were more clearly present in Experiment 2, where subjects were given the most likely hypothesis, than in Experiment 1, where they provided it.

The information heuristic involves failure to consider whether, or with what probability, different test results can lead to different actions. Such failures were found in every experiment in which they could occur (Experiments 1–4). The subject who fails to consider this condition may still consider the first two conditions: (1) whether an answer to a question is consistent with a favored hypothesis, and (2) whether a question can distinguish alternative hypotheses. Use of the first condition (as it is in the congruence heuristic) may favor questions that can raise or lower the probability of the favored hypothesis even when that hypothesis will be

[12] Some of the subjects' justifications in various experiments indicated the use of a reverse congruence heuristic favoring questions that had a high probability of a positive result given some other hypothesis than the most likely. The justifications imply that such a question would serve a kind of self-critical purpose. A similar implication is found in the writing of Wason (1960) and others. In fact, this reverse heuristic is no more self-critical than is the original congruence heuristic. A question that gives a yes answer if the favored hypothesis is true may or may not *discriminate* that hypothesis from an alternative, and the same may be said of a question that gives a yes answer if the alternative hypothesis is true.

TABLE 6
SUMMARY OF THE BIASES AND HEURISTICS FOUND

	Experiment					
	1	2	3	4	5	6
Properties of experiment						
Multiple hypotheses (vs one)	−	−	+	+	+	+
Probabilities given	−	−	−	+	+	+
Joint probability given				−	+	−
Bias or heuristic found						
Congruence	+	+	(+)	−	−	−
Information	+	+	+	+		
Certainty				+	−	+

Note. Plus indicates presence; minus, absence; blank, not applicable. Note that the congruence bias in Experiment 2 was greater than in Experiment 3 (which is therefore in parentheses).

accepted in any case. Use of only the second condition may favor questions that distinguish irrelevant alternatives, which will not be accepted in any case.

In the *certainty* heuristic, subjects value a question that leads to certainty about anything. Experiment 4 revealed a tendency to seek certainty about the presence of the most likely disease (at least). Experiments 5 and 6 indicate that certainty is sought when people have difficulty estimating posterior probabilities. When joint probabilities are provided (in Experiment 5), so that it is easy for subjects to determine which disease is most likely, the use of the certainty heuristic disappears. When people are given conditional probabilities (as in Experiment 6), they apparently would rather seek certainty than solve the problem of how to infer posterior probabilities (or joint probabilities) from the data at hand. Instruction in such calculations (and provision of devices to help people make them) therefore seems to be a likely way to reduce overreliance on the certainty heuristic.

In sum, the congruence and information heuristics may involve failure to carry out different kinds of "checks" on an initial decision to ask a question, hence, a kind of insufficient thinking (Baron, 1985, Chap. 3). The congruence heuristic results from failure to consider alternative hypotheses that might produce the same answer. Provision of a specified alternative hypothesis (as in Experiment 3) makes this check more likely and therefore reduces congruence bias. The information heuristic results from failure to consider relevance for action.

Importantly, many subjects are apparently capable of carrying out the checks that other subjects omit. Many of the subjects who provided justifications made explicit reference to actions that they might take, for

example, "A positive result of test 1 doesn't really tell you much more than what you already suspect, and test 2 is not very prone to give a positive result. I would use test 2 and treat A if + and B if − ." Other subjects in interviews recognized quickly the relevance of conditions they had neglected. The biases reported here may be correctable.

The effect of the biases we have found is undoubtedly most serious in situations in which information is costly, so that is is important to ask exactly the best questions. These situations exist only when it is possible to expend extensive resources to obtain information, as is increasingly the case in science, the professions, and commerce. People who will engage in these activities, such as many of our subjects, may need special training in heuristics for seeking information—either in their chosen fields or in general.

REFERENCES

Allman, R. M., Steinberg, E. P., Keruly, J. C., & Dans, P. E. (1985). Physician tolerance for uncertainty: Use of liver-spleen scans to detect metastases. *Journal of the American Medical Association*, **254**, 246–248.

Baron, J. (1985). *Rationality and intelligence*. New York: Cambridge University Press.

Baron, J., Badgio, P., & Gaskins, I. W. (1986). Cognitive style and its improvement: A normative approach. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, Vol. 3*. Hillsdale, NJ: Erlbaum.

Bursztajn, H., Feinbloom, R. I., Hamm, R. M., & Brodsky, A. (1981). *Medical choices, medical chances: How patients, families, and physicians can cope with uncertainty*. New York: Delacorte Press/Seymour Lawrence.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.

Feldman, M.S., & March, J.G. (1981). Information in organizations as signal and symbol. *Administrative Science Quarterly*, **26**, 171–186.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, **90**, 139–260.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision making under risk. *Econometrica*, **47**, 263–291.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978). Consequences of congruence and disconguence in a simulated research environment. *Quarterly Journal of Experimental Psychology*, **30**, 395–406.

Platt, J. R. (1964). Strong inference. *Science*, **146**, 347–353.

Popper, K. R. (1962). *Conjectures and refutations*. London: Routledge & Kegan Paul.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Shaklee, H., & Fischhoff, B. (1982). Strategies for information search in causal analysis. *Memory and Cognition*, **10**, 520–530.

Tweney, R. D., Doherty, M. E., & Mynatt, C. R. (Eds.). (1981). *On scientific thinking: A reader in the cognitive psychology of science*. New York: Columbia University Press.

Wason, P. C. (1968). On the failure to eliminate hypotheses . . .—A second look. In P. C. Wason & P. N. Johnson-Laird (Eds.), *Thinking and reasoning*. Harmondsworth, Mddx.: Penguin.

RECEIVED: March 10, 1986