Memory & Cognition
1996, 24 (5), 644-654

On people's understanding of the diagnostic implications of probabilistic data

MICHAEL E. DOHERTY, RANDALL CHADWICK, HUGH GARAVAN,
DAVID BARR, and CLIFFORD R. MYNATT
Bowling Green State University, Bowling Green, Ohio

Two lines of prior research into the conditions under which people seek information are examined in light of two statistical definitions of diagnosticity. Five experiments are reported. In two, subjects selected information in order to test a hypothesis. In the remaining three, they selected information in order to convince someone else of the truth of a known hypothesis. A total of 567 university students served as subjects. The two primary conclusions were as follows: (1) When the task is highly structured by the environment, subjects select information diagnostically, and (2) when the task is less structured, so that subjects must seek relevant information not manifest, they select information pseudodiagnostically. Possible relations to other laboratory inference tasks and to clinical judgment are discussed.

In a current text in social psychology, Sabini (1995, citing Trope & Mackie, 1987) has asserted that "an important determinant of whether subjects test hypotheses in a biased way is whether they have a clearly stated alternative in mind. If they do, then subjects tend to be relatively unbiased in their hypothesis testing" (p. 170). This generalization appears to represent a widely held view in the area of social cognition, with variants of it appearing in current textbooks (Brehm & Kassin, 1990; Gilbert, 1995; Smith & Mackie, 1995) and in recent empirical literature (van Wallendael & Guignard, 1992).

The purposes of the present paper are to (1) contrast the research program that supports the above generalization with a research program from outside the literature of social cognition, (2) present some new data, and (3) argue that the above generalization is valid, but only within a relatively restricted domain. First, we note that the concept of diagnosticity is operationalized and defined in at least three different ways, and we focus on two of these. Then we describe the two programs of research, note the essential similarities and differences, and show the respective ties of the two programs to the two definitions of diagnosticity. In five original investigations presented in this paper, we explore some conditions under which people do not show diagnostic data selection.

Diagnosticity

The *diagnosticity of information* may be informally defined as the extent to which that information requires the revision of one's assessment of the probability (P) of some state of the world. Several distinctions must be made for a formal definition of the concept of diagnosticity. We will consider two different ways in which the term has been used, depending on the inferential task involved, with both ways defined in terms of Bayes' theorem. In the odds form, Bayes' theorem is:

$$\frac{P(\mathrm{H}|\mathrm{D})}{P(\sim\mathrm{H}|\mathrm{D})} = \frac{P(\mathrm{H})}{P(\sim\mathrm{H})} \times \frac{P(\mathrm{D}|\mathrm{H})}{P(\mathrm{D}|\sim\mathrm{H})}, \qquad (1)$$

where H and D refer to hypothesis and datum, respectively, and the ~ denotes "not." The most common usage of the term *diagnosticity* refers to the diagnosticity of a datum, that is, the degree to which the occurrence of a datum requires the revision of one's assessment of the probability of H. The *diagnosticity of a datum* is commonly defined in terms of the likelihood ratio (LR), and is derived simply from Equation 1 by segregating the impact of the datum from the prior odds:

$$\mathrm{LR} = \frac{P(\mathrm{D}|\mathrm{H})}{P(\mathrm{D}|\sim\mathrm{H})}. \qquad (2)$$

Another sense in which diagnosticity has been used is with respect to the expected degree to which a question about a feature will, when answered, require the revision of one's assessment of the probability of H. Following Trope and Bassok (1982), we define the *diagnosticity of a question* as the expected likelihood ratio (ELR):

$$\mathrm{ELR} = P(\mathrm{D})\mathrm{LR}(\mathrm{D}) + P(\sim\mathrm{D})\mathrm{LR}(\sim\mathrm{D}), \qquad (3)$$

with the LRs such that LR ≥ 1. The LRs are required to be ≥ 1 since the ELR is designed to reflect the expected change in posterior probability given either answer, irrespective of the direction of change. Hence, Equation 2 reflects the impact of an individual datum, whereas Equation 3 reflects expected diagnostic impact averaged over the possible data. There is also a third sense of the term

This manuscript was prepared with the support of National Science Foundation Grant SBR-9422253 to Bowling Green State University, M.E.D. and C.R.M., principal investigators. The authors would like to acknowledge the contributions to this paper made by Ruth Beyth-Marom, Gernot Kleiter, and an anonymous reviewer. Correspondence should be addressed to: M. E. Doherty, Department of Psychology, Bowling Green State University, Bowling Green, OH 43403 (e-mail: mdoher2@bgnet.bgsu.edu).

diagnosticity that we will not pursue, that is, the diagnosticity of an answer to a question. Whether people who are sensitive to the diagnosticity of a question are sensitive to the potential diagnosticity of each answer is an empirical question we will not address (but see Slowiaczek, Klayman, Sherman, & Skov, 1992).

TWO REFERENCE INVESTIGATIONS

A Study Concluding That People Select Information Diagnostically

Skov and Sherman (1986) presented subjects a scenario in which subjects had to draw an inference about which of two mutually exclusive and exhaustive categories a creature was a member. The scenario dealt with hypothetical creatures on hypothetical planets so that the effects of prior content knowledge would be minimized. Specifically, subjects were asked what questions they would select in order to determine whether an unknown, invisible creature was, for example, a Glom or a Fizo. The subjects "visited" a hypothetical planet on which they selected 2 of 12 available features that discriminated Gloms from Fizos, with diagnosticity defined in terms of the ELR. The conditional probability of each feature, given each category—that is, $P(D|H)$—was also provided. Table 1 presents examples of the features and $P(D|H)$ values. The subjects responded by writing down two *yes/no* questions (e. g., "Do you wear hula hoops?") specifying which two features they would prefer to ask about in order to infer whether the creature was a Glom or a Fizo, with the focal hypothesis having been specified in the instructions. Note that the questions were about features, not about conditional probabilities; if a subject were to ask whether the creature wore hula hoops and that question were to be answered (which it was not), then the subject would know the associated pair of conditional probabilities needed to form a likelihood ratio (LR), since those probabilities were provided along with the features.

The subjects tended strongly to ask the most diagnostic questions, as defined by Equation 3, that is, those designated as 1, 4, 7, and 8 in Table 1. Skov and Sherman (1986) interpreted these results in a straightforward way: "Diagnosticity was the main determinant of question selection. Given a choice between a high diagnostic and hypothesis disconfirming question vs a low diagnostic and hypothesis confirming question, subjects almost always choose the former" (p. 111).

A Study Concluding That People Select Information Pseudodiagnostically

To illustrate this line of research we describe an experiment by Kern and Doherty (1982), who set up a scenario in which subjects had to draw an inference as to which of two fictitious diseases a hypothetical patient had. Fictitious diseases were used to minimize the effects of prior content knowledge. The subjects were 65 advanced medical students who had completed all academic work for the MD degree, and were enrolled in clinical clerkships at a university medical school. The subjects selected the information from a 2×2 array that they considered most useful in making their diagnoses. We quote from a critical part of the instructions:

> You again examine your patient and note that he is running a high fever and is covered with a rash. Having studied the medical history of the island before arriving, you are aware that about an equal number of people on this island suffer from Type A disease as from Type B. (p. 102)

Below that was a 2×2 array of $P(D|H)$ values, where D denotes in this case a symptom, and H denotes disease, but with three of the values covered by opaque stickers. The exposed value revealed that the probability of a rash given Type A disease was .84, and the subject was to select one and only one of the three remaining $P(D|H)$ values in order to make the diagnosis, then write down that diagnosis. The selection was made by peeling off an opaque sticker, revealing the $P(D|H)$ value underneath. In this task, which presents the subjects with two mutually exclusive and exhaustive hypotheses in a static inference problem, the normatively appropriate model of opinion revision, given data, is, as above, Bayes' theorem, but for this paradigm Equation 2 is the relevant one. Given the probability of a rash given Type A disease, the only one of the three remaining conditional probabilities that allows computation of the LR of the symptom is the probability of a rash given Type B disease. That is, if one knows $P(D_1|H_1)$ and can select only one more conditional probability from among $P(D_1|H_2)$, $P(D_2|H_1)$, and $P(D_2|H_2)$, the only $P(D|H)$ that will allow any normatively appropriate computation is $P(D_1|H_2)$. Each subject did two such problems, and of the 65 subjects, only 11 chose the normatively dictated value on both problems. On the basis of these results, Kern and Doherty concluded that "medical students, despite extensive training in patient-information-gathering and decision-making, sought diagnostically irrelevant information when relevant information was equally available" (p. 103). This exemplifies what Doherty, Mynatt, Tweney, and Schiavo (1979) called "the pseudodiagnosticity effect."

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

Feature No.	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

Note—This table is more useful for expository purposes than the corresponding one in Skov and Sherman (1986), since it describes the features semantically as well as statistically. It is taken from Slowiaczek, Klayman, Sherman, and Skov (1992).

A Comparison of the Two Experimental Procedures

These investigations deal with what we believe to be a fundamental cognitive task: how people use probabilis-

tic data to draw inferences. The two investigations are similar in many ways. Both present the subject with information concerning two mutually exclusive and exhaustive categories. Both have the subject gather probabilistic information in order to make an inference, and for both, Bayes's theorem provides the normative standard with which subjects' behavior is compared. Both procedures provide qualitative tests of subjects' understanding of diagnosticity, with neither one requiring numerical responses of the subjects. There are many surface differences, such as subject populations, the specifics of the scenarios, and so on, but the possible influences of these differences are minimized by the variety of investigations using each procedure that lead to the same conclusions.

There are fundamental differences. First, in the investigation by Skov and Sherman (1986), the $P(D|H)$ data were *presented as pairs*. Second, since features rather than $P(D|H)$ values were the object of choice, then, in effect, the $P(D|H)$ values were *chosen as pairs*. The array of features was listed, and the $P(D|H)$ values were explicitly provided. Hence the probative value of each available feature was given to the subjects, and the subjects' task was to select the most diagnostic of the available features and ask whether the creature did or did not have that feature.

Conversely, the subjects in the investigation by Kern and Doherty (1982) were told which features were present, but did not get sufficient information about the diagnosticity of the features that they were given. The subjects were given one $P(D|H)$ value for one feature, and thus had to consider the potential relevance of the remaining data—data that were available but as yet unseen. What is required to produce normative behavior is the insight that the complementary but unseen $P(D|H)$ value that would allow the composition of an LR might be as large or larger than the $P(D|H)$ value provided.

The formal difference between the two tasks is captured by the two meanings of diagnosticity reflected in Equations 2 and 3. The Skov and Sherman (1986) investigation is at the level of the diagnosticity of questions, whereas the Kern and Doherty (1982) investigation is at the level of the diagnosticity of data. In the former, much more structure is provided by the task environment. The latter requires the subjects to impose that structure on the environment. This contrast implies that subjects who have differential diagnosticity displayed for them can at least partially appreciate it, as evidenced by the fact that they tend to make optimal selections of information on which to base their judgments, but that subjects do not appear to have a sufficiently well-developed understanding of the diagnostic implications of data to take into account the possible implications of data that they do not possess. Note that the fact that subjects select diagnostic information appropriately does not mean that they can then use it appropriately, were they to receive it; there is convincing evidence that under certain circumstances they do not (Beyth-Marom, 1990). In order to communicate efficiently, we will refer to tasks in which the pairs of $P(D|H)$ values are displayed or known and the subjects ask which features are present as question diagnosticity (QD) tasks. Tasks in which the features are given and the subjects must seek the $P(D|H)$ values will be called data diagnosticity (DD) tasks.

OTHER INVESTIGATIONS

Investigations Showing Diagnostic Behavior

Several other QD investigations lead to the conclusion that when people are given sets of relevant values—say, $P(D_1|H_1)$ and $P(D_1|H_2)$—they are able to select the best questions wisely. Trope and Bassok (1982) noted that there had been a dearth of research on information gathering in social judgment research, took issue with the widely cited study by Snyder and Swann (1978), and demonstrated that people did select questions according to their diagnosticity. As with the Skov and Sherman (1986) study, the $P(D|H)$ values were an essential aspect of the display presented to subjects. Experiment 2 of Trope and Bassok (1983) involved question selection as the dependent variable, and the subjects evidenced a diagnostic strategy. In that investigation, the $P(D|H)$ values per se were not provided, but verbal descriptions were, so that the probabilistic implications of the verbal descriptions for both hypotheses were manifest. Trope and Mackie (1987) had their subjects assess mutually exclusive and exhaustive hypotheses. In their Experiments 1 and 2, the probabilistic features about which subjects could ask were stated verbally, but they were such that subjects knew the relation between each feature and both hypotheses. As with all of the investigations cited in this section, subjects' question selections (or generations) were systematically diagnostic. Trope and Mackie's Experiment 3 is especially germane to what we posit to be a crucial distinction between these two research paradigms. In Experiment 3, they had subjects formulate questions concerning a category both with and without a specified alternative. Subjects formulated diagnostic questions when the alternative hypothesis was specified, but not when it was unspecified. Kruglanski and Mayseless (1988) also found diagnostic search in a hypothesis-testing task, though their Experiment 2 showed that subjects' search behavior could be influenced by motivational factors (see also Markus & Zajonc, 1985).

Slowiaczek et al. (1992) obtained essentially similar results with respect to question selection (Experiments 3A and 3B) in a series of studies that focused primarily on whether subjects understood the relative usefulness of different possible answers. Van Wallendael and Guignard (1992), in a study also focused primarily on subjects' sensitivity to the probabilistic implications of possible answers, showed that subjects are sensitive to the differential diagnosticity of different questions. And Devine, Hirt, and Gehrke (1990) found question selections to be influenced by diagnosticity as well as by a confirmatory strategy. Finally, Kareev and Halberstadt (1993) showed that people used diagnostic information when it was presented.

Investigations Showing Pseudodiagnostic Behavior

Other investigations suggest that when people are *not* given the relevant $P(D|H)$ values, say $P(D_1|H_1)$ and $P(D_1|H_2)$, they do not seek them. The term *pseudodiagnosticity* was introduced by Doherty et al. (1979), who used a hypothetical archaeological scenario. Like Kern and Doherty (1982)—and unlike the literature just reviewed—Doherty et al. gave subjects the features but had subjects select the conditional probabilities of the features given the hypothesis. There was a very strong tendency for subjects to select $P(D|H)$ values about the same hypothesis, as in Kern and Doherty. A similar effect was obtained in an investigation by Doherty, Schiavo, Tweney, and Mynatt (1981), but that investigation also demonstrated that subjects, with certain sorts of experiences on the task, could learn to behave diagnostically. All subjects were given knowledge of results concerning their inferences, and some were led to see a diagnostic pair of data by having to peel additional stickers. Subjects learned to select $P(D|H)$ values in a normatively appropriate fashion if and only if they were informed that they had made an incorrect inference, and then saw a diagnostic pair.

Other data selection investigations that explicitly show pseudodiagnostic behavior include Beyth-Marom and Fischhoff (1983); Doherty and Mynatt (1990); Mynatt, Doherty, and Dragan (1993); Wolf, Gruppen, and Billi (1985); and Wolf (1983, cited in Wolf et al., 1985). The larger body of evidence referred to as supporting confirmation bias (e.g., Snyder & Swann, 1978) or a positive test bias (e.g., Klayman & Ha, 1989) is relevant, but will not be reviewed here. Wason's (1960, 1968) 4-card and 2–4–6 tasks are also consistent with the general proposition that people are not sufficiently attentive to alternative hypotheses. The picture that emerges from this literature is consistent with Evans's conclusion that people "apply analytic procedures to task features which appear relevant, but they do not actively seek relevant data" (1984, p. 459).

An Investigation Showing Both Diagnostic and Pseudodiagnostic Behavior

We have described what we consider the crucial differences between the groups of investigations that have led to two incompatible conclusions. There are numerous other differences, in addition to the one we have explored: cover scenarios, subject populations, experimenters, specific experimental procedures, and so on. In order to rule these out as explanatory candidates that would render the differences artifactual and uninteresting, Chadwick and Doherty (1993) conducted an experiment in which all of these presumptively extraneous factors were held constant. Subjects were assigned at random to three conditions, with all of the conditions using scenarios that were as similar as possible, consistent with experimental manipulations that would constitute a QD condition, a DD condition, and a novel, mixed-diagnosticity condition that will not be treated further in this paper. The scenario content was taken from Skov and Sherman (1986), described above. We found diagnostic behavior with the QD task and pseudodiagnostic behavior with the DD task. Clearly, the different generalizations under consideration are not the product of accidental aspects of the investigations.

The Present Investigations

The research reviewed above suggests that asking whether subjects understand the diagnostic implications of data is not a good question. There is a better one: Under what conditions are people sensitive to the diagnostic implications of data? The investigations presented below do not follow from one another, as in many research programs. Rather, they are a collection of closely related investigations designed to gain some insight into the conditions influencing the appropriateness of subjects' data selections. The first experiment is a DD study, intended simply to rule out an alternative explanation of the pseudodiagnosticity effect. The second experiment is analogous to a DD study, except that the data are verbal. Experiments 3, 4, and 5 remove the hypothesis-testing aspect of the task in an effort to reduce the demand made on the subjects and vary the task conditions to test the limits of the generalizations about diagnostic and pseudodiagnostic behavior.

EXPERIMENT 1

Experiment 1 is a test of what has been proposed as an artifact in the pseudodiagnosticity paradigm. It has been suggested that people select $P(D_2|H_1)$ rather than $P(D_1|H_2)$ not because of some heuristic such as a positive test strategy, but rather, because they assume that $P(D_1|H_1)$ and $P(D_1|H_2)$ sum to 1.00—or, since we present these as percentages, to 100%. Given that assumption, it would be reasonable for subjects to select $P(D_2|H_1)$ given $P(D_1|H_1)$. This alternative explanation has been raised several times, most recently by Lopes (personal communication, 1991). Hence, in this investigation, we gave subjects highly salient instructions that the $P(D|H)$ values could sum to any value.

Method

Subjects. The subjects were 155 introductory psychology students at Bowling Green State University. They received credit toward a course participation requirement.

Materials and Procedure. There were three different content problems, one involving a diagnosis, one involving an inference about which political party a speaker represented, and the third involving the identification of the street on which a friend lived. Each problem was in two forms, one with the focal hypothesis mentioned in the final sentence of the scenario as well as being indicated by a visible $P(D|H)$, and the other with the focal hypothesis indicated only by a visible $P(D|H)$.

Each subject received a two-page booklet. Unlike previous DD investigations, preceding the scenario on which subjects made their data selections there was an instruction page that presented a sample inference problem with all four $P(D|H)$ values in plain sight, arranged in a 2×2 array. The $P(D|H)$ values were such that the

rows summed to 131% and 84%, and the columns to 117% and 98%. The subjects had merely to inspect the problem; neither data selections nor inferences were called for. Immediately below the data array was the following sentence, in boldface italics: "Notice that there is no need for the percentages in either rows or columns to sum to 100%. All 4 percentages might be 100%; all 4 might be 0% or anything in between." These instructions also convey to sophisticated subjects that the data are conditionally independent, thus ruling out another possible criticism, though one would expect conditional nonindependence to lead subjects to behave diagnostically, rather than pseudodiagnostically. The experimental DD problem was on the second page, one version of which follows:

> Your friend has moved to a new house and you're trying to remember where it is. You've narrowed it down to the houses on Street X and houses on Street Y. You know that he lives in a brick house that is worth over $100,000 in value. One more piece of information is shown, telling you that 81% of the houses on Street X are worth over $100,000.
>
> Which other piece of information from the table below would be most helpful in figuring out which street your friend lives on?

	Street X	Street Y
Percentage of houses that are over $100,000	81%	57%
Percentage of houses made of brick	65%	45%

> Please pick *only one piece of information* from the remaining three in the problem. We are very interested in which piece of information you think will be most useful to you. Take your time in making this choice and in answering the question below.

Only the 81% was exposed, the other three being covered by opaque stickers. Subjects were asked to identify the street on which their friend lived, and to mark their degree of belief on a $P(H|D)$ scale that ranged from .50 to 1.0. The other two problems had $P(D_1|H_1)$, $P(D_1|H_2)$, $P(D_2|H_1)$, and $P(D_2|H_2)$ values of 77%, 57%, 61%, and 45%, and 79%, 57%, 40%, and 45%, respectively, with $P(D_1|H_1)$ being the exposed value in all cases.

Results and Discussion

Neither the manipulation of the salience of the focal hypothesis nor the content of the scenarios made a systematic difference in data selection or in the judged $P(H|D)$ responses at the bottom of the page. Hence, for simplicity, the data will be pooled over the six experimental conditions. Of the 155 subjects, 45 selected $P(D_1|H_2)$, the diagnostic response, 69 selected $P(D_2|H_1)$, the pseudodiagnostic and positive test bias response, and 41 selected $P(D_2|H_2)$, a response explained by subjects in other studies as wanting to get some information about each datum and each hypothesis. Note that the selection of $P(D_2|H_2)$ is pseudodiagnostic, but would not qualify as a +H test. A test of goodness-of-fit showed that subjects were not simply selecting data randomly [$\chi^2(2, N = 155) = 8.87$, $p < .05$]. The mean $P(H|D)$ response for all subjects drawing the correct inference was virtually identical for the three response categories: .683, .685, and .665, for $P(D_1|H_2)$, $P(D_2|H_1)$, and $P(D_2|H_2)$ data selections, respectively. Only 11 subjects drew erroneous conclusions, with 6 of those having made the $P(D_2|H_2)$ selection.

Experiment 1 had a limited purpose: to see if the pseudodiagnosticity effect would obtain under conditions in which it would be difficult to argue that the effect was due to a misinterpretation of the independence of the conditional probabilities. There was a substantial degree of pseudodiagnostic data gathering. The level of diagnostic behavior was higher than in most of our studies, so the assumption that $P(D|H)$ and $P(D|{\sim}H)$ sum to 1, which we might refer to as "illicit complementation," may have been one source of variance in data selection behavior in earlier DD tasks. Nevertheless, there is strong evidence that the pseudodiagnosticity effect is a real one; the subjects selected inappropriate data and confidently based their inference thereon. The fact that the subjects who selected pseudodiagnostic data drew the correct inference as often as those subjects who selected diagnostic data does not suggest that the strategy is an ecologically useful one; their drawing of the correct inference is due to the specific $P(D|H)$ values in the tables. The study could as easily have been designed to mislead the subjects. Whether the pseudodiagnostic strategy is a good one outside the laboratory would require what Brunswik (1956) called an ecological survey.

EXPERIMENT 2

Experiment 2 was conducted to assess the generalizability of the tendency to seek more information about the hypothesis about which one already has information. All information was presented in nonnumerical form; conditional probabilities were not given. In the absence of quantitative information, there is no normatively correct Bayesian data selection strategy. Thus for the present problems, the issue of what data selection choices are diagnostic or pseudodiagnostic is moot. There is, on other grounds that will be described below, a best data selection.

Method

Subjects. Subjects were 96 introductory psychology students. Participation in the experiment partially fulfilled a class requirement.

Materials. Each subject was given a booklet containing 16 problems, 8 of which are relevant to this paper. Each problem began with a brief statement that provided a context for the data selection followed by two mutually exclusive alternatives and two mutually exclusive categories of information. Each problem can be conceptualized as a 2 × 2 matrix, as above, but subjects were presented data choices in the form of statements. Let a statement of the form, "Hypothesis 1 is characterized by Level 1 of Feature 1," be called Cell A, a verbal analog of $P(D_1|H_1)$. Similarly, let the analogs of $P(D_1|H_2)$, $P(D_2|H_1)$, and $P(D_2|H_2)$ be called Cells B, C, and D, respectively, as in the example below.

One datum (Cell A) was presented in each problem stem. Data in the other three cells were presented below each problem in the form of three phrases. One problem is shown below. The cell designations are shown for purposes of exposition only, and were not on the versions seen by subjects.

> You are taking a botany class and have been given the assignment of deciding whether a particular tree is a Canadian Birch or an American Birch. You are first told that the tree bears leaves which are reddish in color and oval in shape.
>
> You are then told that a Canadian Birch has oval-shaped leaves (Cell A) and you must make your judgment based on only one more item of information. Which will you choose?
>
> A. The shape of leaves from American Birch. (Cell B)
>
> B. The color of leaves from Canadian Birch. (Cell C)
>
> C. The color of leaves from American Birch. (Cell D)

The initial datum given in the problem was randomly selected for each problem from among the four possible information category-

alternative relationships. The order in which the information equivalent to the three remaining cells was presented was balanced so that each order appeared an equal number of times for each problem across all subjects and on each page of the booklet. This was accomplished by creating six versions of each problem, varying only in the order of the three phrases. Each subject's booklet was arranged in a different random order, thus partially controlling for possible carryover effects from one judgment to the next.

The type of alternative was also manipulated. For half of the problems, both of the decision alternatives were specific. For example, in the tree problem, subjects were told that an unknown tree was either a Canadian birch or an American birch. In the other half of the problems, one of the decision alternatives was specific and the other was nonspecific. For example, in the fish problem, subjects were told that an unknown fish was either a frostfish or some other kind of fish. This manipulation was included primarily in light of Beyth-Marom and Fischhoff's (1983) speculations about the effect of specificity. Although there was no effect of this variable in their study, we felt that it was of sufficient interest to manipulate it in the present study.

Instructions and Procedure. Subjects were run in groups of 8 to 30 but worked on the booklets individually. They were told that the experiment was concerned with judgment strategies and that they would find a description of a problem requiring a judgment on each page of the booklet. They were instructed to read each problem carefully, to consider the information presented in the description, and then to circle the letter next to the one additional piece of information they would select from the three presented below the description.

Results and Discussion

Table 2 shows the choice data for all eight problems, broken down by cell. On each of the eight problems, 1×2 χ^2 tests of goodness-of-fit indicated that subjects made significantly more Cell C choices than Cell B choices. Cell D was rarely chosen. The ratio of Cell C to Cell B choices is almost 5:1. As in Beyth-Marom and Fischhoff (1983), the specificity of the alternative hypothesis had no effect on the choice frequencies. Hence these results are consistent with the general finding that subjects prefer more data about the same hypothesis over data relevant to the alternative.

Table 2
Frequency of Cell B, C, and D Choices for Each Problem in Experiment 2

	Cell Choices		
Problem	B	C	D
	Specific Alternatives		
Disease	0	96	0
Tree	20	69	7
Tooth	17	73	6
Wreck	35	55	6
Total	72	293	19
	Nonspecific Alternatives		
Fish	10	80	6
Bar	9	87	0
Oil	2	92	2
Suspect	30	55	11
Total	51	314	19

Note—For all eight comparisons, Cell C frequencies are significantly greater than Cell B frequencies [$\chi^2(1, N \geq 85) \geq 7.35, p < .01$].

This experiment considered alone, however, admits of a ready alternative explanation to that just proposed. The results are consistent with the generalization that people seek more information about the hypothesis about which they already have positive information, but they are also consistent with the proposition that the subjects select the most informative data. Given the reasonable assumption that some datum will differentiate between the two possible hypotheses, a subject who requests information from either Cell B or Cell D may be left uncertain about the correct classification, depending on the answer. No matter what the answer to Cell C, however, the identity of the tree is determined. Hence, unlike the probabilistic form, it is optimal in this case to seek more information about the same hypothesis. The fact that in this case the strategy is a good one does not diminish the interest of the data or the power of the generalization; there may be other task environments in which the tendency to seek more data about the same hypothesis is optimal, as in, for example, a number of the conditions in Mynatt, Doherty, and Sullivan (1991), and as explored in Klayman and Ha (1987).

We have indicated what we considered the essential distinction between QD and DD tasks, and alluded to Evans's (1984) theorizing that subjects simply failed to see potential data concerning H_2 as relevant to conclusions about H_1. It is possible that the uncertainty inherent in having to consider both the possible truth and falsity of the hypothesis in question results in cognitive overload, preventing an otherwise functional understanding of diagnosticity to come to the fore. Hence, in Experiments 3, 4, and 5, we reduced the task uncertainty for the subjects.

In the following three experiments, we sought to assess whether or not subjects, *when they knew the true hypothesis*—that is, the identity of the creature, in addition to knowing the features characterizing the creature—would show a tendency to select information diagnostically in order to support the truth of that hypothesis. We sought to reduce the cognitive load by eliminating the hypothesis—testing part of the task. Subjects in Experiment 5, in addition to knowing the true hypothesis and the features, knew the $P(D|H)$ values associated with each feature.

The cover stories in Experiments 3 and 4 were similar to the one in Skov and Sherman's (1986) investigations of hypothesis testing. In the present studies, however, subjects were told, "You have met one of these creatures and . . . it is a Glom," and were instructed to select information that they would provide to someone else to prove that the creature was indeed a Glom. In these hypothesis-supporting tasks, subjects selected individual conditional probabilities that could form zero, one, or two LRs, and normative data selection is still dictated by Bayes's theorem.

GENERAL METHOD, EXPERIMENTS 3–5

Subjects. A total of 316 subjects from introductory psychology classes participated in these three experiments.

Table 3
Number of Subjects Forming Zero, One, or Two Likelihood Ratios (LRs) in Experiments 3, 4, and 5

	Experiment					
	3					
No. LRs	Glom	Fizo	Neutral	4	5	Total
0	42	42	33	41	32	190
1	10	10	13	6	16	55
2	11	8	20	18	14	71
Total	63	60	66	65	62	316

Note—"Forming" indicates the number of pairs of data selected that would form LRs. In each task, a maximum of two LRs could be formed.

Materials and Procedure. Each subject completed one task, presented on a single sheet. In all versions, the subject met one of two hypothetical creatures on a hypothetical planet, was told that the creature was a Glom, and that there were equal numbers of Gloms and Fizos on the planet.

Experiments 3, 4, and 5 were run simultaneously, with participants randomly assigned to tasks. The experimenters handed out task sheets to subjects, read general verbal instructions from a script, and told the participants to begin. When a sufficient amount of time had elapsed for all subjects to complete their tasks, approximately 20 min, the experimenters gathered the sheets and debriefed the subjects.

EXPERIMENT 3

In Experiment 3, we examined data selection strategies when participants were presented with two mutually exclusive and exhaustive hypotheses, told which hypothesis was true, and then asked to select $P(D|H)$ values supporting the true hypothesis. We investigated the extent to which providing two conditional probabilities needed to form one LR might alter information selection strategies.

Method

Subjects. Subjects were 189 introductory psychology students. Participation in the experiment partially fulfilled a class requirement.

Materials and Procedure. Subjects received one of three versions of a DD task, in which they were shown a 2 × 5 data matrix consisting of five features together with the probabilities of each feature, under the two hypotheses—that is, $P(D_1|H_1)$, $P(D_1|H_2)$... $P(D_5|H_1)$, $P(D_5|H_2)$. Of these 10 $P(D|H)$ values, 3 were visible; both $P(D|H)$ values for Feature 1 were shown, as was $P(D_2|H_1)$, which was .92 for all conditions. Each of the other seven $P(D|H)$ values was covered by an opaque sticker. Subjects were instructed to select any three $P(D|H)$ values by peeling three stickers.

There were three conditions. In the Glom condition the conditional probabilities associated with Feature 1 were $P(D|Glom) = .85$ and $P(D|Fizo) = .35$, for an LR of 2.43 favoring the Glom (true) hypothesis. A second (Fizo) condition reversed these same conditional probabilities to favor the Fizo hypothesis. A third condition (neutral) had conditional probabilities of $P(D|Glom) = .85$ paired with $P(D|Fizo) = .88$, for an LR of 1.04, which obviously did not strongly favor either hypothesis. It was expected that under the neutral condition, information selection would be most diagnostic, under the assumption that the exposed conditional probabilities would show that a high $P(D|H)$ could be paired with an equally high $P(D|\sim H)$. In other words, the independence of the numerator and denominator of the LR was made salient.

Subjects were randomly assigned to the Glom, Fizo, or neutral condition (60, 63, and 66 subjects, respectively). They were instructed in writing to select the three $P(D|H)$ values that would best convince a colleague that the creature was a Glom, and to do so by peeling the stickers from any three of the covered conditional probabilities.

Results and Discussion

Table 3 shows the number of subjects choosing data in such a way as to form zero, one, or two LRs. A χ^2 was done on each of the three conditions to test the null hypothesis that sticker selection was random. For all three conditions, $\chi^2(2, N \geq 60) \geq 80.61$, $p < .001$; hence subjects were selecting data in a systematic fashion. However, only 13% of the subjects in the Fizo condition, 17% in the Glom condition, and 30% in the neutral condition behaved normatively, that is, according to Bayes's rule. In all three conditions, the most common data selection pattern was to select all three $P(D|H)$ values from the same hypothesis as that for which the $P(D|H)$ value had been given. These results conform to previous results using the DD task and suggest that reduction of cognitive load does not affect data selection strategies.

To see if showing $P(D|H)$ values with an LR ≈ 1.0 influenced data selection strategies, subjects in the Glom and Fizo conditions, in which the LRs departed substantially from 1.0, were combined and compared with the neutral condition subjects. Proportionately twice as many neutral subjects behaved normatively as those who had LRs that were large [$\chi^2(1, N = 189) = 5.79$, $p < .05$]. Note again, however, that a minority of subjects in all conditions chose data according to the normative model. Given that the two exposed $P(D|H)$ values associated with Feature 1 summed to greater than 1, these data support the conclusion of Experiment 1 that illicit complementation cannot explain pseudodiagnosticity.

EXPERIMENT 4

In Experiment 4, we assessed subjects' patterns of data selection when actual numerical values of conditional probabilities were not provided. It might be argued that our relatively naive subjects were biased somehow by the numerical $P(D|H)$ values, and in this experiment we sought to determine what change, if any, would occur in the absence of such potential biases. Persistence of pseudodiagnostic behavior in the face of this change would buttress the conclusion of Experiment 2, which would be desirable given the alternative explanation of the results of that experiment.

Method

Subjects. Subjects were 65 introductory psychology students. Participation in the experiment partially fulfilled a class requirement.

Materials and Procedure. The task was similar to that used in Experiment 3; subjects were given the true hypothesis (Glom) and were asked to select evidence that favored this hypothesis. In this version, however, no numerical data were given and subjects were to imagine that by selecting a datum they would have the associated $P(D|H)$ value. The eight features (e.g., "The percentage of Gloms who eat iron ore") were arranged in one column, each followed by a blank. An X was marked in that blank next to $P(D_1|Glom)$, indicating that that $P(D|H)$ was known. Subjects were to select three more data by placing an X in three other blanks.

Results and Discussion

The frequencies of subjects choosing zero, one, or two LRs are shown in Table 3. Subjects were not selecting data randomly [$\chi^2(2, N = 65) = 107.18, p < .001$]. Of the 65 subjects, only 18 (28%) selected data normatively. This percentage may be artifactually high due to the fact that the columns were arranged so that if a subject were simply to place an X in the three blanks immediately below that which represented $P(D_1|H_1)$, data selection would appear normative. Nevertheless, the results are in line with other pseudodiagnosticity studies.

EXPERIMENT 5

In Experiment 5, the data were presented exactly as in Skov and Sherman (1986), but the subjects were free to choose individual $P(D|H)$ values (as opposed to features that automatically provided LRs). Subjects were presented with four pairs of $P(D|H)$ values from which they were to choose the four individual values that would best convince a skeptical colleague that the creature they had met was indeed a Glom. The cover story elaborated reasons why having seen all the $P(D|H)$ values was not the same as being able to bring those $P(D|H)$ values back. Note that this task is similar in format to many of the tasks used in the QD research. $P(D|H)$ values were presented as pairs, with each pair associated with a particular feature.

There were two key differences between the present investigation and typical QD studies, one in the cover story and one in the nature of the response. In Experiment 5, the subjects were informed not only of the identity of the creature, but also of its specific characteristics. That is, subjects were informed that it was a Glom, that it eats iron ore, drinks gasoline, has blue skin, and wears a hula hoop. The difference in the response was that our subjects selected *individual* $P(D|H)$ values to bring back to their skeptical colleague, rather than being constrained to select pairs of $P(D|H)$ values.

The task allows for normative behavior, in that subjects may select data that would form two LRs. Unlike the typical QD task, however, this variant of the task also allows for nonnormative behavior, in that subjects may select $P(D|H)$ values that do not enable the formation of any LRs. Hence in the present investigation, subjects must have sufficient understanding of the principle of diagnosticity to impose the necessary organization on the data.

Method

Subjects. Subjects were 62 introductory psychology students. Participation in the experiment partially fulfilled a class requirement.

Materials and Procedure. The task was a modified version of the cover story employed by Skov and Sherman (1986), with the instructions altered as described above. The $P(D|H)$ pairs—that is, $P(D|Glom)/P(D|Fizo)$—were presented as 85%/35%, 92%/68%, 41%/7%, and 75%/68%, though not, of course, displayed as ratios.

Results and Discussion

There was a clear preference for data describing the known creature, with 172 conditional probability values for Gloms selected in contrast to 76 for Fizos. Of the subjects who showed a preference, 40 selected more $P(D|Glom)$ values, and just one selected more $P(D|Fizo)$ values.

Diagnostic behavior is reflected in the selection of conditional probabilities that allow for the construction of LRs. For ease of communication we will refer to this as the selection of diagnostic pairs, or just pairs, though subjects selected not pairs but individual conditional probabilities. Normative behavior, that is, selecting $P(D|H)$ values such that two LRs could be formed, was evidenced in 14 subjects; 16 subjects chose one pair and 32 chose no pairs [$\chi^2(2, N = 62) = 29.16, p < .001$].

A closer examination of the pairs chosen proves interesting. Of all pairs chosen (44 in total), the most diagnostic one was selected 19 times, the second most diagnostic 20 times, the third 5 times, and the least diagnostic pair not at all. Those pairs that subjects did select reveal a distinct tendency for people who choose pairs to choose diagnostic ones over those of low diagnosticity, which is what the QD research has shown.

The modal selection pattern (selected by 18 subjects) was the 7% $P(D|H)$ for Fizos and the three $P(D|H)$s for Gloms, with no paired $P(D|H)$ values chosen. This is further evidence that subjects believe that a single $P(D|H)$ is informative, which has also been shown by Beyth-Marom (1990).

As noted, one purpose of this block of experiments was to see if subjects' strategies would be more frequently normative if the task uncertainty were reduced. The failure of the reduction of task uncertainty to reduce pseudodiagnostic data selection can be interpreted in at least three ways. First, it is hard to know how much the subjects' cognitive loads were actually reduced. Second, as discussed in Mynatt et al. (1993), subjects have a tendency to seek further information about the hypothesis they believe to be true. A third possibility, closely related to the second, is that subjects, upon seeing a conditional probability, think that the complementary conditional probability is irrelevant. This is essentially the meaning of the term *pseudodiagnosticity*—that is, that a $P(D|H)$ value alone has diagnostic value.

GENERAL DISCUSSION

These results demonstrate clearly that the quotation with which we opened this paper is an overgeneralization; one cannot make general statements about the degree to which subjects seek information diagnostically without qualification in terms of task conditions. There is solid evidence in prior research that subjects show diagnostic behavior when selecting questions (with diagnosticity defined in terms of Equation 3), but show pseudodiagnostic behavior when selecting data needed to form LRs (with diagnosticity defined in terms of Equation 2). Taken together, prior research and the present experiments show that subjects will select data normatively if and only if certain situational constraints make salient the relevance of data about H_2 to inferences about H_1; the $P(D|H)$ values must be presented as pairs and the subjects must se-

lect them as pairs. That is, if the situation constrains the subjects' attention to focus directly on the relation between the values in the stimulus array, the subjects will very likely select information diagnostically. However, if subjects must bring that degree of structure to the information themselves, it is unlikely that they will do so. The data of Experiment 5 speak especially clearly to this point; most subjects chose data pseudodiagnostically, but those subjects who did choose pairs of $P(D|H)$ values were likely to choose them in a diagnostic fashion.

These investigations also show that one source of variance in the pseudodiagnosticity effect is the extent to which the task forces subjects to attend to the idea that $P(D_i|H_1)$ and $P(D_i|H_2)$ do not normally sum to 1.0. But even then, the number of subjects choosing information diagnostically is small, considering that optimal performance would mean that 100% of subjects do so. Given the tasks used, whether the information to be selected is verbal or numerical had no influence on pseudodiagnostic information selection, but we certainly would not generalize that null finding to other task domains. Finally, the results show that information selection is not influenced by whether the subject knows the truth of the hypothesis about which evidence is sought. There are, of course, other task parameters that influence information selection (e.g., the extremity bias shown by Skov & Sherman, 1986) not addressed in this paper.

Theoretical Implications

We believe that the distinction drawn above is directly related to Evans's conception of *heuristic*, which refers to "*pre-attentive* processes whose function is to select relevant information for analytic processing" (Evans, 1984, p. 452). The QD and DD research paradigms assess subjects' data selection for analytic processing. The two paradigms lead to different generalizations because in one, QD, the displays and the requirement to select features defined by pairs of $P(D|H)$ values make the relevance of diagnostic pairs transparent. In the DD task, subjects appear to make a preanalytic judgment of relevance that includes only data that relate to the hypothesis they have reason to believe true. That is, they simply fail to see potential data concerning H_2 as relevant to conclusions about H_1. Evans referred to heuristic processes as "rapid and indescribable," a set of descriptors reminiscent of Brunswik's (1956) distinction between perception and thinking, and Hammond's distinction between intuition and analysis (Hammond, Hamm, Grassia, & Pearson, 1987). Broadly, the conclusion is that subjects have a sufficient understanding of question diagnosticity to select the most diagnostic questions when the environment structures the task, as in the QD paradigm, but not a sufficient understanding of the diagnosticity of the data, as in the DD paradigm, to seek out relevant information that is not presented. In addition, Beyth-Marom (1990) and Slowiaczek et al. (1992) have shown that people are not sensitive to the diagnosticity of answers in the domains tested.

We also construe this set of findings as consistent with the three propositions concerning data selection posited in Mynatt et al. (1993):

1. *People will normally test hypotheses that they believe true, rather than hypotheses that they believe false.* It is well established that people tend to have difficulty with negation (Wason & Johnson-Laird, 1972) and have a generalized positivity bias (Evans, 1989). With respect to the present investigations, this proposition is reflected in Experiments 1, 3, 4, and 5.

2. *The number of objects that can be maintained and operated on in working memory is one.* An object can be a hypothesis, a dimension of utility, or an explicit relation between two of these, depending on the task. We construe the QD results as situations in which the object is the relation between $P(D|H)$ and $P(D|\sim H)$.

3. *People commonly update their beliefs on the basis of information relevant to the single hypothesis in working memory.* This proposition is reflected especially clearly in Experiment 5 and in Beyth-Marom (1990; see also Robinson & Hastie, 1985, and van Wallendael & Hastie, 1990).

This set of assumptions is compatible with traditional conceptions of memory, which hold that there is an active memory, or short-term store, a subset of which is the "focus of attention" (Cowan, 1988, 1993).

Relation to Other Laboratory Tasks

The failure to attend to the potential relevance of information about alternative hypotheses is also reflected in other laboratory investigations. For example, in reviewing the literature on Wason's 4-card selection task, Tweney and Doherty (1983) noted that the notorious difficulty with that task might be due to the failure of subjects to consider the relevance of the unchosen cards to the possibility that the rule might be false. This is related to the proposition (see, e.g., Evans, 1989, p. 60; Klayman & Ha, 1987) that subjects tend to focus on positive information. One clear finding in the research using Wason's 2–4–6 task is that subjects do not readily generate triples to test alternative hypotheses (Wason, 1960), unless the experimental situation is designed to get them to do so (Gorman, Stafford, & Gorman, 1987; Tweney et al., 1980; Wharton, Cheng, & Wickens, 1993). It is our experience with the 2–4–6 task that subjects simply *never* spontaneously state two hypotheses when performing a test with a single triple (Tweney et al., 1980), which is consistent with the proposition that subjects typically consider data as relevant to only one hypothesis at a time. Nor do subjects in our artificial universe studies state multiple hypotheses (Garavan, 1992; Mynatt, Doherty, & Tweney, 1978).

Another related phenomenon that is readily observed in the laboratory is illusory correlation, one manifestation of which is the tendency to report relationships when none exists (Arkes, 1981; Chapman & Chapman, 1969; Crocker, 1981). The DD and QD experimental paradigms bear a close formal relation to the typical illusory correlation paradigm. Consider a common illusory correlation task, the inference of whether there is a relation between a

symptom and some illness, based on a number of observations of each of the four events in the 2 × 2 table constructed from data concerning the presence or absence of that symptom and that illness. The pairs of $P(D|H)$ and $P(D|\sim H)$ that are the crux of the DD and QD research are simply two of the marginal probabilities of that 2 × 2 table. It is commonly believed that a major contributor to the illusory correlation effect is the tendency of subjects to discount as irrelevant all cause-absent data (Arkes, 1981). This strategy leaves the subject with the information needed to construct $P(D|H)$ but not with the information needed to construct $P(D|\sim H)$, which is consistent with our explanation of the pseudodiagnosticity effect.

Perhaps the research most closely related to the DD and QD research is that of Klayman and Brown (1993) and McKenzie (1994). Their results are highly consistent with the DD and QD results, despite considerable differences in the research paradigms; their subjects learned the relations between data and hypotheses via case-by-case exposure to frequentistic information. Some learned the relations between symptoms and each of two possible illnesses one at a time; others between symptoms and both of two illnesses at the same time. The latter procedure was called "contrastive learning." Those authors theorized that the former learning protocol would lead to "independent representations" of each illness, so that subjects would consider the implications of symptoms for a single disease (see also Van Wallendael, 1989). On the other hand, contrastive learning was postulated to lead to a dependent representation, such that evidence relevant to H_1 would be seen as relevant to H_2. The difference between independent and dependent representations corresponds to what we see as how subjects react to DD and QD task structures, respectively, and the experimental results are very much in accord with each other.

Borrowing Beyth-Marom and Fischhoff's (1983) words, "If one wished to summarize these data into a statement about people's abilities as intuitive statisticians, one would need to specify the context within which their abilities were being tested. If information is selected and organized for them, then they generally show a qualitative understanding of diagnosticity. Unfortunately, however, organized presentations are probably the exception rather than the rule in everyday experience" (p. 1194).

Implications for Inference Beyond the Laboratory

We believe that we are talking about basic psychological processes when we assert the three propositions concerning data selection for hypothesis testing. We restrict our speculation about the implications of these processes to two important categories of inference in the world, employment interviewing and clinical psychodiagnosis. It has long been the received view among investigators in the area of industrial/organizational psychology that the categorical decision (hire/don't hire) is made in the very first few minutes of the interview session, and that the often considerable time spent thereafter is devoted to the search for corroborating information (Webster, 1982). In the terms noted above, once H_1 (suitability, or its converse) is adopted, the search for data relevant to $\sim H_1$ is abandoned.

While this pseudodiagnostic data selection *cum* +test strategy may have immediate negative consequences for the interviewee and long-term negative consequences for the organization, this form of data selection may have terribly negative consequences in the clinical domain. Dawes (1994) explored the consequences of judgment biases, among other issues, for clinical practice. We believe that the general tendency toward pseudodiagnosticity *cum* +test strategy is at least in part at the root of a problem he so powerfully decries. The issue is important; we quote at length:

> Now consider the statement that "I can identify child abusers because I have had experience working with 50 [or 100, or even 500] of them." Child abuse may have a fairly precise definition on the basis of actual behavior, but professionals who attempt to learn from experience to distinguish abusers from nonabusers must—according to the learning from errors principle—have experience with people who *appear* to be child abusers but are *not*. Where does such information come from? It is extraordinarily difficult to obtain; in fact, it is impossible to obtain if one's contact is limited to people who actually are child abusers." (Dawes, 1994, p. 119)

The relation to pseudodiagnosticity is clear; a clinician may have an idea of how frequently a feature (symptom) may be associated with a psychiatric condition—that is, $P(D|H)$—and base a diagnosis in part on that information without considering how frequently that feature may occur in the absence of that category—that is, $P(D|\sim H)$. This example highlights the intimate relation between the failure to attend adequately to $P(D|\sim H)$ and the phenomenon of base rate neglect (Bar-Hillel, 1990; Kahneman & Tversky, in press; Koehler, 1996). The situation for the clinician is compounded by two traditions of psychodiagnosis. The first is the virtually exclusive tendency to focus on symptoms of psychopathology; one searches the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM–IV*, 1994) of the American Psychiatric Association in vain for a description of the "symptoms" of health. The second is the reliance on verbal rather than statistical formulations; one searches the *DSM–IV* in vain for representations of $P(D|H)$ and $P(D|\sim H)$.

In many circumstances, actions depend upon inferences. Clearly, we believe that the failure to seek information that might favor hypotheses other than the hypothesis that is at the moment the focus of one's attention may lead to erroneous inferences, and as a result, nonoptimal actions.

REFERENCES

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting & Clinical Psychology*, **49**, 323-330.

Bar-Hillel, M. (1990). Back to base rates. In R. M. Hogarth (Ed.), *Insights in decision making* (pp. 200-216). Chicago: University of Chicago Press.

Beyth-Marom, R. (1990). Mis/understanding diagnosticity: Direction and magnitude of change. In K. Borcherding, O. I. Larichev, & D. M.

Messick (Eds.), *Contemporary issues in decision making* (pp. 203-221). Amsterdam: Elsevier, North-Holland.

BEYTH-MAROM, R., & FISCHHOFF, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality & Social Psychology*, **45**, 1185-1195.

BREHM, S. S., & KASSIN, S. M. (1990). *Social psychology*. Boston: Houghton Mifflin.

BRUNSWIK, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley: University of California Press.

CHADWICK, R., & DOHERTY, M. E. (1993, November). *Inattention to data relevant to alternative hypotheses*. Paper presented at the annual meeting of the Psychonomic Society, Washington, DC.

CHAPMAN, L. J., & CHAPMAN, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, **74**, 271-280.

COWAN, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin*, **104**, 163-191.

COWAN, N. (1993). Activation, attention, and short-term memory. *Memory & Cognition*, **21**, 162-167.

CROCKER, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin*, **90**, 272-292.

DAWES, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth*. New York: Free Press.

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. (1990). Inattention to P(H) and to P(D/~H): A converging operation. *Acta Psychologica*, **75**, 1-11.

DOHERTY, M. E., MYNATT, C. R., TWENEY, R. D., & SCHIAVO, M. D. (1979). Pseudodiagnosticity. *Acta Psychologica*, **43**, 111-121.

DOHERTY, M. E., SCHIAVO, M. B., TWENEY, R. D., & MYNATT, C. R. (1981). The influence of feedback and diagnostic data on pseudodiagnosticity. *Bulletin of the Psychonomic Society*, **18**, 191-194.

EVANS, J. ST. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology*, **75**, 451-468.

EVANS, J. ST. B. T. (1989). *Bias in human reasoning: Causes and consequences*. London: Erlbaum.

GARAVAN, H. (1992). *When falsification fails*. Unpublished master's thesis, Bowling Green State University.

GILBERT, D. T. (1995). Attribution and interpersonal perception. In A. Tesser (Ed.), *Advanced social psychology* (pp. 99-147). New York: McGraw-Hill.

GORMAN, M. E., STAFFORD, A., & GORMAN, M. E. (1987). Disconfirmation and dual hypotheses on a more difficult version of Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology*, **39A**, 1-28.

HAMMOND, K. R., HAMM, R. M., GRASSIA, J., & PEARSON, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on Systems, Man, & Cybernetics*, **SMC-17**, 753-770.

KAHNEMAN, D., & TVERSKY, A. (in press). On the reality of cognitive illusions: A reply to Gigerenzer's critique. *Psychological Review*.

KAREEV, Y., & HALBERSTADT, N. (1993). Evaluating negative tests and refutations in a rule discovery task. *Quarterly Journal of Experimental Psychology*, **46A**, 715-727.

KERN, L., & DOHERTY, M. E. (1982). "Pseudodiagnosticity" in an idealized medical problem-solving environment. *Journal of Medical Education*, **57**, 100-104.

KLAYMAN, J., & BROWN, K. (1993). Debias the environment instead of the judge: An alternative approach to reducing error in diagnostic (and other) judgment. *Cognition*, **49**, 97-122.

KLAYMAN, J., & HA, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, **94**, 211-228.

KLAYMAN, J., & HA, Y.-W. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **15**, 596-604.

KOEHLER, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative and methodological challenges. *Behavioral & Brain Sciences*, **19**, 1-53.

KRUGLANSKI, A. W., & MAYSELESS, O. (1988). Contextual effects in hypothesis testing: The role of competing alternatives and epistemic motivations. *Social Cognition*, **6**, 1-20.

MARKUS, H., & ZAJONC, R. B. (1985). The cognitive perspective in social psychology. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology: Vol. 1. Theory and method* (3rd. ed., pp. 137-230). New York: Random House.

MCKENZIE, C. R. M. (1994). *Taking into account the strength of an alternative hypothesis*. Unpublished doctoral dissertation, University of Chicago.

MYNATT, C. R., DOHERTY, M. E., & DRAGAN, W. (1993). Information relevance, working memory, and the consideration of alternatives. *The Quarterly Journal of Experimental Psychology*, **46A**, 759-778.

MYNATT, C. R., DOHERTY, M. E., & SULLIVAN, J. A. (1991). Data selection in a minimal hypothesis testing task. *Acta Psychologica*, **76**, 293-305.

MYNATT, C. R., DOHERTY, M. E., & TWENEY, R. D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology*, **30**, 395-406.

ROBINSON, L. B., & HASTIE, R. (1985). Revision of opinion when a hypothesis is eliminated from consideration. *Journal of Experimental Psychology: Human Perception & Performance*, **11**, 443-456.

SABINI, J. (1995). *Social psychology*. New York: Norton.

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.

SLOWIACZEK, L. M., KLAYMAN, J., SHERMAN, S. J., & SKOV, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition*, **20**, 392-405.

SMITH, E. R., & MACKIE, D. M. (1995). *Social psychology*. Worth.

SNYDER, M., & SWANN, W. B. (1978). Behavioral confirmation in social interaction: From social perception to social reality. *Journal of Experimental Social Psychology*, **14**, 148-162.

TROPE, Y., & BASSOK, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality & Social Psychology*, **43**, 22-34.

TROPE, Y., & BASSOK, M. (1983). Information-gathering strategies in hypothesis-testing. *Journal of Experimental Social Psychology*, **19**, 560-576.

TROPE, Y., & MACKIE, D. M. (1987). Sensitivity to alternatives in social hypothesis-testing. *Journal of Experimental Social Psychology*, **23**, 445-459.

TWENEY, R. D., & DOHERTY, M. E. (1983). Rationality and the psychology of inference. *Synthèse*, **57**, 139-161.

TWENEY, R. D., DOHERTY, M. E., WORNER, W., PLISKE, D., MYNATT, C. R., GROSS, K., & ARKKELIN, D. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology*, **32**, 109-123.

VAN WALLENDAEL, L. R. (1989). The quest for limits on noncomplementarity in opinion revision. *Organizational Behavior & Human Decision Processes*, **43**, 385-405.

VAN WALLENDAEL, L. R., & GUIGNARD, Y. (1992). Diagnosticity, confidence, and the need for information. *Behavioral Decision Making*, **5**, 25-37.

VAN WALLENDAEL, L. R., & HASTIE, R. (1990). Tracing the footsteps of Sherlock Holmes: Cognitive representations of hypothesis testing. *Memory & Cognition*, **18**, 240-250.

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. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, **20**, 273-281.

WASON, P. C., & JOHNSON-LAIRD, P. N. (1972). *Psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University Press.

WEBSTER, E. C. (1982). *The employment interview: A social judgment process*. Schomberg, ON: S.I.P. Publications.

WHARTON, C. M., CHENG, P. W., & WICKENS, T. D. (1993). Hypothesis-testing strategies: Why two goals are better than one. *Quarterly Journal of Experimental Psychology*, **46A**, 743-758.

WOLF, F. M., GRUPPEN, L. D., & BILLI, J. E. (1985). Differential diagnosis and the competing-hypothesis heuristic. *Journal of the American Medical Association*, **253**, 2858-2862.

(Manuscript received June 6, 1994;
revision accepted for publication August 2, 1995.)