Psychological Review
1983, Vol. 90, No. 3, 239-260

Hypothesis Evaluation From a Bayesian Perspective

Baruch Fischhoff
Decision Research, Perceptronics, Inc.
Eugene, Oregon and
Medical Research Council Applied Psychology Unit
Cambridge, England

Ruth Beyth-Marom
Decision Research, Perceptronics, Inc.
Eugene, Oregon

Bayesian inference provides a general framework for evaluating hypotheses. It is a normative method in the sense of prescribing how hypotheses should be evaluated. However, it may also be used descriptively by characterizing people's actual hypothesis-evaluation behavior in terms of its consistency with or departures from the model. Such a characterization may facilitate the development of psychological accounts of how that behavior is produced. This article explores the potential of Bayesian inference as a theoretical framework for describing how people evaluate hypotheses. First, it identifies a set of logically possible forms of nonBayesian behavior. Second, it reviews existing research in a variety of areas to see whether these possibilities are ever realized. The analysis shows that in some situations several apparently distinct phenomena are usefully viewed as special cases of the same kind of behavior, whereas in other situations previous investigations have conferred a common label (e.g., *confirmation bias*) to several distinct phenomena. It also calls into question a number of attributions of judgmental bias, suggesting that in some cases the bias is different than what has previously been claimed, whereas in others there may be no bias at all.

Hypothesis evaluation is a crucial intellectual activity. Not surprisingly, it is also a focus of psychological research. A variety of methods have been applied to understand how people gather and interpret information in order to evaluate hypotheses. Either implicitly or explicitly, some theory of how people *should* evaluate hypotheses provides the conceptual framework for studies of how they *do* evaluate them. Such a prescriptive theory provides a set of articulated terms for describing tasks and a definition of appropriate behavior against which actual performance can be compared. The theory might even be descriptively valid at a certain level if people are found to follow its dictates, either due to natural predilections or because they have been trained to do so. Even when behavior is suboptimal, some psychological insight may be obtained by asking whether that behavior may be described as a systematic deviation from the theory. Reference to a normative theory can also identify performance deficits that need to be understood and rectified.

One general and popular normative scheme is Bayesian inference, a set of procedures based upon Bayes' theorem and the subjectivist interpretation of probability. These procedures show how to (a) identify the data sources that are most useful for discriminating between competing hypotheses, (b) assess the implications of an observed datum vis-à-vis the truth of competing hypotheses, (c) aggregate the implications of different data into an overall appraisal of the relative likelihood of those hypotheses being correct, and

Support for this research came from the U.S. Office of Naval Research under Contract N00014-80-C-0150 to Perceptronics, Inc. and from the United Kingdom Medical Research Council.

Our thanks go to Hal Arkes, Jack Dowie, Ward Edwards, Sarah Lichtenstein, Nils-Eric Sahlin, Paul Slovic, Amos Tversky, and an anonymous reviewer for thoughtful comments on earlier drafts of this article.

Requests for reprints should be sent to Baruch Fischhoff, Decision Research, A Branch of Perceptronics, 1201 Oak, Eugene, Oregon 97401. (Ruth Beyth-Marom is now at the Open University, P.O. Box 39328, Ramat Aviv 61392 Israel.)

(d) use that appraisal to select the course of action that seems best in light of available evidence. Excellent detailed expositions of the scheme may be found in Edwards, Lindman, and Savage (1963), Lindley (1965), Novick and Jackson (1974), and Phillips (1973).

The present article presents a simple version of Bayesian inference. From this scheme, it derives a taxonomy of logically possible deviations. This taxonomy is then used to characterize several published studies reporting biased hypothesis evaluation. In some cases, the result is to reiterate the claims of the original investigators; in other cases, those claims are countered by alternative interpretations that suggest other biases that may be involved or ways in which observed behavior might be construed as being properly Bayesian. In still other cases, research conducted in other traditions is cast in Bayesian terms in the hope of profiting from others' experience and drawing different fields together.

This review is not meant to be exhaustive; rather, it is meant to illustrate how the Bayesian perspective can be used to illuminate a variety of tasks. As a result, it emphasizes new interpretations and, by reference to the taxonomy, it identifies potential biases for which positive evidence is lacking—raising the question of whether it was an opportunity to observe suboptimal behavior that researchers missed or an opportunity to exhibit suboptimal behavior that subjects "missed." The theory is complete in the sense that it treats all of the basic issues that arise in Bayesian inference and that must be faced when one assesses the optimality of behavior. It is incomplete in that it does not show how the theory can be adapted to model all possible situations. For example, all of the hypotheses we consider here are discrete, both for simplicity's sake and because the studies we cite have used discrete hypotheses. The treatment of continuous hypotheses (e.g., "Today's mean temperature is x") is considered by Edwards et al. (1963) and others.

Theory

Definition of Probability

From the Bayesian perspective, knowledge is represented in terms of statements or hypotheses, H_i, each of which is characterized by a subjective probability, $P(H_i)$, representing one's confidence in its truth (DeFinetti, 1976). For example, one might be .75 confident that it will snow tomorrow. Such probabilities are subjective in the sense that individuals with different knowledge (or beliefs) may legitimately assess $P(H_i)$ quite differently. The term *assess* is used rather than *estimate* to emphasize that a probability expresses one's own feelings rather than an appraisal of a property of the physical world. Thus, there is no "right" probability value for a particular statement. Even if a very low probability proves to be associated with a true statement, one cannot be sure that it was not an accurate reflection of the assessor's (apparently erroneous) store of knowledge.

The constraints on subjective probabilities emerge when one considers sets of assessments. Formally speaking, a set of probabilities should be orderly or *coherent* in the sense of following the probability axioms (Kyburg & Smokler, 1980; Savage, 1954).[1] For example, $P(H) + P(\bar{H})$ should total 1.0.

Updating

Additional derivations from the probability axioms lead to Bayes' theorem, which governs the way in which one's beliefs in hypotheses should be updated in the light of new information. In its simplest form, the theorem treats the implications of an observation that produces the datum (D) for determining whether a hypothesis (H) is true, relative to its complement, $\bar{H}$. In such cases, Bayes' theorem states that

$$\frac{P(H/D)}{P(\bar{H}/D)} = \frac{P(D/H)}{P(D/\bar{H})} \cdot \frac{P(H)}{P(\bar{H})}. \qquad (1)$$

Reading from the right, the three terms in this formula are (a) the prior odds that H (and not $\bar{H}$) is true in the light of all that is known before the receipt of D, (b) the likelihood ratio, representing the information value of D with respect to the truth of H, and (c) the posterior odds that H is true in the light of all that is known after the receipt of

[1] The demand of coherence is what differentiates DeMorgan's "pure subjective" interpretation of probability—as whatever people actually believe—from DeFinetti's personalistic interpretation of probability as rational belief (see Kyburg & Smokler, 1980).

D. Equation 1 could also be applied to any pair of competing hypotheses, A and B (with A replacing H and B replacing $\bar{H}$).

Although the subjectivist interpretation of probability is controversial, Bayes' theorem generally is not. The axioms from which it is derived are common to most interpretations of probability. Bayesians have more frequent recourse to the theorem because the subjectivist position enables them to incorporate prior beliefs explicitly in their inferential processes.[2]

Likelihood Ratios

If the probability of observing D given that H is true is different from the probability of observing D when H is not true, then the likelihood ratio is different from 1 and the posterior odds are different from the prior odds. That is, the odds favoring H have become smaller or greater as a result of having observed D. Such a datum is considered to be informative or *diagnostic*. Its degree of diagnosticity can be expressed in terms of how different the likelihood ratio is from 1. Clearly, diagnosticity depends upon the hypotheses being tested. A datum that distinguishes one hypothesis from its complement may be completely uninformative about another pair of hypotheses. Data do not answer all questions equally well.

The value of the likelihood ratio is independent of the value of the prior odds. One could in principle observe a datum strongly supporting a hypothesis that is initially very unlikely. If that happened, one's posterior odds favoring H might still be very low, although not as low as they were before. There is also no necessary relationship between the values of the numerator and the denominator of the likelihood ratio. A datum might strongly favor H even if it is very unlikely given the truth of H. Similarly, observing a datum that is a necessary concomitant of H, that is, $P(D/H) \approx 1$, may be uninformative if it is also a necessary concomitant of $\bar{H}$.

Action

The apparatus of Bayesian inference also provides tools for converting one's beliefs in hypotheses into guides to action. In simplest terms, these tools translate the cost associated with erroneously acting as though a hypothesis is true and the cost of erroneously acting as though a hypothesis is false into a *critical ratio*. If the posterior odds favoring H are above this value, then one is better off acting as though H is true; if they are below the value, then one is better off acting as though H is false.

The value of the critical ratio depends, of course, on the particular kinds of action that are contemplated and the outcomes that are associated with them. Where one's posterior odds stand vis-à-vis the critical ratio depends on both one's prior odds and the evidence that has subsequently been received. Two individuals who had agreed on the costs of the different errors and on the meaning of the different data might still act differently, if they had different prior odds. On the other hand, if the cumulative weight of the new evidence was sufficient, they might act in the same way, despite having initially had quite discrepant beliefs—for example, if the evidence carried them from prior odds of 1:10 and 1:1 to posterior odds of 100:1 and 1,000:1, respectively.

When it is possible to collect more data, additional Bayesian procedures can help identify the most useful source (Raiffa, 1968). Such *value-of-information* analyses evaluate the expected impact of each possible observation on the expected utility of the actions one can take. They can also tell when the cost of further observation is greater than its expected benefit.

Ways to Stray

Bayesian inference, like other normative schemes, regards only those who adhere to its every detail as behaving optimally. Conversely, every component of the scheme offers some opportunity for suboptimality. This section catalogues possible pitfalls. For the normative minded, these possibilities might be seen as defining Bayesian inference negatively by pointing to behavior that is inconsistent with it.

[2] Although beyond the scope of the present article, discussion of cases in which Bayes' theorem might not be the most useful way of updating beliefs may be found in Diaconis and Zabell (1982), Good (1950), and Jeffrey (1965, 1968). One troublesome situation is receiving information that changes one's whole system of beliefs [and not just $P(H)$].

For the descriptive minded, these logical possibilities suggest judgmental biases that might be observed in empirical studies. If observed in situations in which people are properly instructed and motivated to respond correctly, such deviations can be theoretically informative because they seem to reflect deep-seated judgmental processes. From a practical standpoint, such deviations suggest opportunities for constructive interventions that might lead to better inferences and, subsequently, to better decisions based on those inferences. These interventions might include training, the use of decision aids, or the replacement of intuition by formalized procedures (Edwards, 1968; Fischhoff, 1982; Kahneman & Tversky, 1979).

These potential biases are presented in Table 1. The left-hand column shows the task in which the problem could arise; the center column describes the possible biases; and the right-hand column points to phenomena reported in the literature that we have interpreted as special cases of these biases, which are described in detail in the following section.

Hypothesis Formation

For hypothesis evaluation to begin, there must be hypotheses to evaluate. Indeed, because the diagnostic impact of data is defined only in the context of particular hypotheses, there is no systematic way that data can even be collected without such a context. In the absence of any hypotheses, the collection of data is merely idle stockpiling. Although it is a logical possibility, it does not seem to be a troublesome or common bias. In practice, even the most ambling data collection may be guided by a vague idea of the hypotheses that the collector might be asked to evaluate. Or, it may be conducted for exploratory purposes, with the goal of generating, rather than evaluating, hypotheses. If the collector's goal is to see something that fortuitously stimulates a creative insight, then the Bayesian model, or any other formal model, can offer little guidance or reprobation.

A more serious threat than the absence of hypotheses to evaluate is the possession of hypotheses that cannot be evaluated. One route to formulating hypotheses that cannot

Table 1
Potential Sources of Bias in Bayesian Hypothesis Evaluation

Task	Potential bias	Special cases
Hypothesis formation	Untestable	Ambiguity, complexity, evidence unobservable
	Nonpartition	Nonexclusive, nonexhaustive
Assessing component probabilities	Misrepresentation	Strategic responses, nonproper scoring rules
	Incoherence	Noncomplementarity, disorganized knowledge
	Miscalibration	Overconfidence
	Nonconformity	Reliance on availability or representativeness
	Objectivism	—
Assessing prior odds	Poor survey of background	Incomplete, selective
	Failure to assess	Base-rate fallacy
Assessing likelihood ratio	Failure to assess	Noncausal, "knew-it-all-along"
	Distortion by prior beliefs	Preconceptions, lack of convergence
	Neglect of alternative hypotheses	Pseudodiagnosticity, inertia, cold readings
Aggregation	Wrong rule	Averaging, conservatism?
	Misapplying right rule	Computational error, conservatism?
Information search	Failure to search	Premature conviction
	Nondiagnostic questions	Tradition, habit
	Inefficient search	Failure to ask potentially falsifying questions
	Unrepresentative sampling	—
Action	Incomplete analysis	Neglecting consequences, unstable values
	Forgetting critical value	Confusing actual and effective certitude

be evaluated lies through ambiguity, either intentional or inadvertent. To take a popular example, astrology columns offer hypotheses about what consequences follow what acts (e.g., "You will be better off avoiding risky enterprises"). Yet these acts and consequences are so vaguely defined that it is unclear whether what eventually happens supports the hypothesis. A more sophisticated form of ambiguity can be found in the probabilistic risk analyses used to generate detailed hypotheses about the operation of technical systems (e.g., "Toxins can be released to the atmosphere only if the following events occur"; Green & Bourne, 1972; U.S. Nuclear Regulatory Commission, Note 1). Although actual operating experience should afford an opportunity to evaluate these hypotheses, it may be unclear whether the specific events that are observed are subsumed by the generic events described in the analysis. For example, investigators disagreed over whether the fire at the Browns Ferry Nuclear Power Plant in 1975 was included among the accident sequences described in the then-definitive analysis of reactor operation (U.S. Nuclear Regulatory Commission, Note 2, Note 3).

Complexity offers a second route to formulating hypotheses that cannot be evaluated. Hypotheses that are set out clearly may have such great detail that no datum provides a clear message. For example, political advisors may escape charges of having predicted events incorrectly by noting that every last detail of their advice was not followed ("Had they only listened to me and done X and Y as well, then everything would have been all right").[3] O'Leary, Coplin, Shapiro, and Dean (1974) found that among practitioners of international relations (i.e., those working for business or government) theories are of "such complexity that no single quantitative work could even begin to test their validity" (p. 228). Indeed, some historians argue that the accounts of events that they produce are not hypotheses at all but attempts to integrate available knowledge into a coherent whole. In this light, a valid explanation accommodates all facts, leaving none to test it. This attitude toward hypotheses has its own strengths and weaknesses (Fischhoff, 1980).

Because it considers the relative support that evidence provides to competing hypotheses, the Bayesian scheme requires not only the individual hypotheses but also the set of hypotheses to be well formulated. In effect, those hypotheses must partition some space of possibilities.[4] Computationally, problems arise with nonexclusive hypotheses, which render the message of evidence ambiguous.

Whereas achieving mutually exclusive hypotheses requires precision of formulation, securing an exhaustive set of hypotheses often requires the exercise of imagination. Technically, it is easy to make a set exhaustive by defining it in terms of H and $\bar{H}$, by declaring the hypotheses that one has thought of to be the only ones of interest, or by adding a hypothesis consisting of "all other possibilities," a role that $\bar{H}$ may fill. Although such specification is adequate for some purposes, wherever the alternatives to H are poorly defined, it is hard to evaluate $P(D/\bar{H})$ and, hence, the implications of D for H. Whenever the set of alternatives is incomplete, major surprises are possible. The difficulties of exhaustive enumeration are often exploited by mystery writers who unearth a neglected hypothesis that tidily accounts for available evidence. Both for generating hypotheses and for evaluating $P(D/\text{all other possibilities})$, an important skill is estimating the completeness of the set of already listed hypotheses. Evidence suggests that people tend to exaggerate the completeness of hypothesis sets (Fischhoff, Slovic, & Lichtenstein, 1978;

[3] A topical example may be found in the runaway inflation that has followed the linking of incomes and loans to the cost-of-living index in several countries. Although his economic theories predicted the opposite result, Milton Friedman has denied that this unhappy experience constitutes evidence against his theories because the countries involved did not implement the indexation exactly as he prescribed. Even more troubling for the status of his theories about the economy is that further thought (perhaps stimulated by this irrelevant experience) has led him to conclude that his earlier derivation was wrong and that, in fact, indexation encourages inflation under some conditions ("How Indexation Builds In Inflation," 1979).

[4] In the case of multiple hypotheses, the posterior probability $P(H_i/D) = P(D/H_i)P(H_i)/\sum_i P(D/H_i)P(H_i)$. A diagnostic datum is then one for which $P(D/H_i) \neq \sum_i P(D/H_i)P(H_i)$.

Mehle, Gettys, Manning, Baca, & Fisher, 1981).

A final problem is that even well-formulated hypotheses may be wrong for the actions contemplated. The acid test of relevance is whether perfect information about the hypothesis (i.e., knowing whether it is true or false) would make any difference to one's actions. For example, if ethical principles proscribe incarcerating juveniles with adults, then hypotheses about the effect of prison on delinquents have no consequences for criminal policy. If their support for military spending depends upon the relative strength of the pro- and anti-arms lobbies, then legislators need not choose among competing hypotheses regarding Soviet intentions. The pejorative label for such impractical hypotheses is "academic." More charitably, they are "hypotheses relevant to possible future actions."

Assessing Component Probabilities

To find a place in the Bayesian model, one's beliefs must be translated into subjective probabilities of the form appearing in the model. Any difficulties in assessing such component probabilities would impair hypothesis evaluation. As mentioned, the Bayesian perspective holds probabilities to be subjective expressions, reflecting the assessor's beliefs. Accepting the subjectivist position does not, however, mean accepting any probability as an appropriate assessment of someone's state of belief. There are a number of ways in which the assessment of probabilities can go wrong.

One possible problem is lack of candor. People may misstate their beliefs, perhaps to give a response that is expected of them, perhaps to avoid admitting that an unpleasant event is likely to happen, or perhaps to achieve some strategic advantage by misrepresenting how much they know or what they believe. When the truth of statements can eventually be ascertained, the use of proper scoring rules should encourage candor (Murphy, 1972). These rules reward people as a function of both their stated beliefs and the (eventually revealed) state of the world in such a way that the probability with the highest expected value is the one expressing one's true belief. Whether these rules prove effective in practice is a moot point (Lichtenstein, Fischhoff, & Phillips, 1982; von Winterfeldt & Edwards, 1982). Where they do not prove effective or where they cannot even be applied because the truth will not be known or because no reward system is possible, other means of assuring candor are needed.

The crucial assessment problem from the subjectivist perspective is lack of coherence, failure of one's assessments to follow the probability axioms. Violations may be due to a poor job of reviewing one's knowledge. For example, $P(H)$ and $P(\bar{H})$ may not equal 1 when the two hypotheses are not evaluated simultaneously—and concentration on each evokes a different subset of one's knowledge. It is also possible that the beliefs themselves are not well thought through. In that case, working harder on probability assessment may lead to even more incoherent probabilities by revealing the underlying inconsistencies. Here, the structure of belief requires refinement (Lindley, Tversky, & Brown, 1979).

Another symptom of poor assessment is *miscalibration,* failure of one's confidence to correspond to reality. If a hypothesis with a prior probability of .2 eventually is found to be true, the initial assessment is dubious. However, it may be an accurate summary of the assessor's knowledge at that moment. Recurrent association of high prior probabilities with hypotheses that prove to be false should be cause for serious concern. Calibration formalizes this reliance on the eventually accepted truth of hypotheses for validating probability assessments. Specifically, for the well-calibrated assessor, probability judgments of .*XX* are associated with true hypotheses *XX*% of the time. Empirical studies of calibration have shown that probabilities are related, but are not identical, to proportions of correct hypotheses. The most common deviation is overconfidence—for example, being only 80% correct when 1.00 confident (Lichtenstein, Fischhoff, & Phillips, 1982).[5] As with incoherence, miscali-

[5] Those who interpret subjective probabilities in terms of intuitively appropriate betting odds would never say 1.00 because that would mean willingness to bet everything on the truth of the hypothesis involved. In that light, the only reasonable interpretation of 1.00 is as "nearly 1.00" or "above .995 and rounded upward."

bration may be traced to people's assessment procedures or to the knowledge base that they rely on.

A third symptom is *nonconformity,* producing a probability that differs from that of "expert" assessors for no apparent reason. The existence of such consensus is most likely when there is a statistical data base upon which to base probability assessments (e.g., public health records of mortality). In this restricted realm, the distinction between subjective and objective probabilities becomes blurred, as subjectivists would typically act as though they concur with the relative-frequency interpretation of probability, which objectivists consider to be the only meaningful one. Subjectivists would, however, never concede that frequency counts are a completely objective measure of probability, arguing that judgment is needed to establish the equivalence and independence of the set of trials from which the count was extracted and to extrapolate that frequency to future events.

Another way to look at bias is in terms of the process by which assessments are produced. There is reason for concern whenever the assessors have followed procedures that are inconsistent with the rules of statistical inference and are unaware of those inconsistencies. Two well-known deviations are reliance on the availability and representativeness heuristics when making probability assessments (Kahneman, Slovic, & Tversky, 1982; Tversky & Kahneman, 1974). Users of the former judge an event to be likely to the extent that exemplars are easy to recall or imagine; users of the latter judge an event to be likely to the extent that it represents the salient features of the process that might produce it. Although both rules can provide good service, they can also lead the user astray in predictable ways. For example, reliance on availability induces overestimation of unusually salient events (e.g., the probability of dying from flashy, hence overreported, causes such as tornadoes and homicide; Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978).

A final process problem is the refusal to consider anything but relative frequency data when one assesses probabilities. Although subjectivists acknowledge the potential relevance of such data, they will not be bound to them. Indeed, a key selling point of Bayesian inference is its ability to accommodate diverse kinds of data. One can, for example, assess probabilities for a meaningful shrug or an off-hand comment as well as for a bead drawn from an urn or for a 40-subject experiment. The only difference is that as one moves from beads to shrugs, it becomes increasingly difficult to attest to the reasonableness of a particular assessment. An assessor who failed to seek useful and available nonfrequency evidence would, from a Bayesian perspective, be foolish. An individual who ignored nonfrequency evidence that was already on hand would be biased.[6]

Assessing the Prior

Prior beliefs are captured by the ratio of the probabilities of the competing hypotheses prior to the collection of further evidence. As a result, the difficulties of assessing prior beliefs are by and large the sum of the difficulties of formulating hypotheses and assessing component probabilities, as already discussed. There seem, however, to be at least two additional (and incompatible) problems that may affect this particular stage of the assessment process.

One such problem is not treating the component probabilities equally. An extreme form of inequality is neglecting one of the hypotheses. When people act as though a hypothesis that is most probably true is absolutely true, they have effectively neglected its complement. In that case, hypothesis evaluation never begins, because one hypothesis is treated as fact. Even when both hypotheses are considered, one may be given deferential treatment. Skeptics, for example, may give undue weight to evidence that contradicts a favored hypothesis; they have a warm spot in their hearts for complements. On the other hand, Koriat, Lichtenstein, and Fischhoff (1980) found favoritism for initially favored

[6] One difficulty that the Bayesian approach avoids is vagueness in expressing beliefs. People often disagree considerably about the interpretation of verbal expressions of likelihood (e.g., "probably true"). Moreover, the same individual may use a term differently in different contexts (Beyth-Marom, 1982; Reyna, 1981). The Bayesian approach requires explicitness.

hypotheses: Subjects who were asked to determine the relative likelihood of two possible answers to a question seemed to search primarily for supporting evidence.[7] Such directed search may serve legitimate purposes, for example, seeing if any case at all can be made for, or against, a particular hypothesis. However, it may be very difficult to estimate and to correct for the bias that it induces in the resultant sample of evidence. Indeed, it is the failure to realize the biases in the samples produced by the availability heuristic that makes it a potentially dangerous judgment rule.

These difficulties vanish in the presence of another bias that has attracted considerable attention of late: neglecting the base rate (Kahneman & Tversky, 1973). The *base-rate fallacy* refers to the tendency to allow one's posterior beliefs to be dominated by the information that one extracts from the additional datum, *D*, to the neglect of the prior beliefs, expressing what typically has been observed. The several recent reviews of this literature (Bar-Hillel, 1980; Bar-Hillel & Fischhoff, 1981; Borgida & Brekke, 1981) may be summarized as indicating that people rely most heavily on whatever information seems most relevant to the task at hand. Thus, for example, when testing a pair of hypotheses such as "John is a lawyer/John is an engineer," even weak diagnostic information relating directly to John may dominate base-rate information reporting the overall prevalence of the two professions. Base-rate information is used, however, if it can be linked more directly to the inference or if the case-specific information is palpably worthless.

The evidence for the base-rate fallacy comes primarily from studies in which base-rate information was presented, yet neglected. As such, the fallacy can be viewed as a problem of aggregation, which we treat two sections below. It is discussed here because the failure to use explicitly presented base rates strongly suggests that they will not be spontaneously sought or assessed. This suspicion is confirmed, for example, by Lyon and Slovic's (1976) finding that, when asked directly, one half of their subjects did not believe that base rates were relevant to their judgments.[8]

Assessing the Likelihood Ratio

In principle, people can ignore the likelihood ratio just as well as the base rate, allowing one to speak of the "likelihood-ratio fallacy" whenever people fail even to consider the likelihood ratio for a pertinent datum. This may happen, for example, when the datum provides merely circumstantial evidence (Einhorn & Hogarth, 1982), when it cannot be woven into a causal account involving the hypothesis (Tversky & Kahneman, 1980), or when it reports a nonoccurrence. A classic example of the latter is Sherl●ck Holmes's observation (Doyle, 1974) that his colleague, Inspector Gregory, had not considered the significance of a dog failing to bark when an intruder approached.

Failure to assess the likelihood ratio of received evidence may also be encouraged by hindsight bias, underestimating the informativeness of new data (Fischhoff, 1975, 1982). The feeling that one knew all along that *D* was true might make the calculation of *D*'s likelihood ratio seem unnecessary. Denying that *D* has anything to add does not, however, mean that it will not have any impact—only that one will be unaware of that impact. At the same time as people deny its contribution, new information can change their thinking in ways that they cannot appreciate or undo (Fischhoff, 1977; Sue, Smith, & Caldwell, 1973). These unintended influences may or may not be those that would follow from a deliberative analysis.

[7] One might argue that the collection of diverse pieces of existing evidence should not be considered the *assessment of a base rate,* a term that should be reserved for aggregating statistical data (e.g., 70% of previous cases have supported *H*). Following that argument would lead one to the Laplacian assumption that all hypotheses are equally likely a priori except in the presence of statistical data to the contrary. In that case, the problems discussed in the text would be relegated to the section on assessing likelihood ratios associated with the diverse data. A further argument holds that even statistical data were separate pieces until they were aggregated—meaning that their aggregation required the use of likelihood ratios. At this extreme, a priori odds would always be equal to 1.

[8] Ascribing neglect to subjects requires confidence that the investigator knows what base rates really are relevant on the basis of what subjects have been told. Although this determination is relatively clear in most experiments, it can be quite difficult in real-life problems.

When judges do assess the likelihood ratio, the sequencing of that operation may expose it to the influence of the preceding operations. In particular, the interpretation of new evidence may be affected by previous beliefs, thereby subverting the independence of the likelihood ratio and priors. Nisbett and Ross (1980) offer an impressive catalog of ways in which people can interpret and reinterpret new information so as to render it consistent with their prior beliefs. So great is people's ability to exploit the ambiguities in evidence to the advantage of their preconceptions and to discount inconsistent evidence, that erroneous beliefs can persevere long after they should have been abandoned.[9]

Such biased interpretation of evidence also thwarts the effective convergence of belief that should follow use of Bayesian inference. However discrepant the initial position of two individuals, their posterior beliefs should converge for practical purposes, providing they observe a sufficiently large set of diagnostic data about whose interpretation they agree. The ability to interpret a datum as supporting contradictory hypotheses means that convergence may never occur. Whatever data the two observers see, they become more convinced of their respective prior beliefs.

When people choose to evaluate evidence, they must compare two conditional probabilities, $P(D/H)$ and $P(D/\bar{H})$. Such comparison is essential because there is no necessary relationship between these two components of the likelihood ratio. A variety of studies suggest, however, that people consider only the numerator. That is, they are interested in how consistent the evidence is with the hypothesis they are testing, $P(D/H)$ and fail to consider its consistency with the alternative hypothesis, $P(D/\bar{H})$. As a result, the size of $P(D/H)$ determines D's support for H. Users of this strategy act as though they assume that the two conditional probabilities are inversely related, although, in principle, both may be high or low. A datum with a low $P(D/H)$ may provide strong evidence for H if $P(D/\bar{H})$ is even lower; a datum for which $P(D/H)$ is high may reveal nothing if $P(D/\bar{H})$ is equally high.

Four examples should give some flavor of the variety of guises within which incomplete appraisal of the likelihood ratio may emerge:

1. Doherty, Mynatt, Tweney, and Schiavo (1979) presented subjects with six pieces of data, D_i and allowed them to inquire about the values of 6 of the 12 conditional probabilities: $P(D_1/H)$, $P(D_1/\bar{H})$, . . . , $P(D_6/H)$, $P(D_6/\bar{H})$. Few subjects requested any of the pairs of conditional probabilities [e.g., $P(D_3/H)$ and $P(D_3/\bar{H})$]needed to compute likelihood ratios. The authors labeled this tendency to pick but one member of each pair, *pseudodiagnosticity.* The probabilities that subjects did solicit were primarily those, $P(D/H_i)$, involving the hypothesis, H, that a preceding manipulation had made appear more likely. The authors called this tendency *confirmatory bias* (a term to which we will return).

2. Troutman and Shanteau (1977) had subjects draw beads from a box which contained either 70 red, 30 white, and 50 blue beads or 30 red, 70 white, and 50 blue beads and infer the probability that the box was predominantly white (W). Drawing two blue beads reduced subjects' confidence in their initially favored hypothesis (W), even though that observation is equally unlikely under either hypothesis: $P(D/R) = P(D/W) = .11$. Thus, subjects considered only $P(D/H)$ for their favored hypothesis.

3. In a similar experiment, without the blue beads, Pitz, Downing, and Reinhold (1967) found that subjects who were fairly confident that the box was predominantly red would increase their confidence in that hypothesis after observing a white. Subjects apparently felt that they should see an occasional white, neglecting the fact that that event was still more likely if the box being used was predominantly white. Pitz et al. called this failure of inconsistent evidence to slow the increase of confidence in H an *inertia effect.* The inappropriate increase in faith in H here contrasts with the inappro-

[9] The undue influence of prior information in this context is in sharp contrast to the neglect of prior information observed with the base-rate fallacy. One difference between the two cases is that in the former case the prior beliefs are actually posterior odds that subjects had generated by actively weighing previous evidence. In many base-rate fallacy studies, the prior beliefs were only a statistical summary reporting what someone else had typically observed. A second difference is that the perseverating prior beliefs were more specific than the neglected ones, fitting Bar-Hillel's (1980) account.

priate decrease in it in a study by Troutman and Shanteau (1977). We would trace these two opposite effects to the same underlying cause, neglecting $P(D/\bar{H})$.

4. A favorite ploy of magicians, mentalists, and pseudopsychics who claim to read other people's minds is to provide universally valid personality descriptions (Forer, 1949; Hyman, 1977) that apply to almost everyone, although this is not transparently so. These operators trust their listeners to assess *P*(this description/my mind is being read) and not *P*(this description/my mind is not being read).

A final threat to the validity of assessed likelihood ratios comes from the fact that the probabilities involved all concern conditional events. This added complexity seems to complicate probability assessment, with people forgetting the conditioning event, reversing the roles of the two events, or just feeling confused (Eddy, 1982; Moskowitz & Sarin, in press).

Aggregation

Assuming that judges have attended to and assessed all components of the Bayesian model, they must still combine them to arrive at posterior odds. The two logically possible sources of error here are (a) using the wrong aggregation rule, for example, averaging, rather than multiplying, the likelihood ratio and prior odds, and (b) using the right rule, but applying it inappropriately, for example, making a computational error. Establishing whether either or both of these potential biases actually occurs was a focus of early research into intuitive Bayesian inference (reviewed by Slovic & Lichtenstein, 1971).

Most of these early studies used the un-Bayesian strategy of creating inferential tasks in which the experimenter could claim to know the correct subjective probability for all participants. This was done by using highly artificial stimuli for which all reasonable observers should have the same subjective probability. For example, subjects might be shown a series of poker chips and be asked to evaluate the hypotheses: They are being drawn from a bookbag with 70% blue chips and 30% red chips/they are being drawn from a bag with 70% red chips and 30% blue chips. The predominant result of this research was that subjects' confidence in the apparently correct hypothesis did not increase as quickly as the accumulating evidence indicated that it should.

A lively debate ensued over whether this poor performance, called *conservatism*, reflected failure to appreciate how diagnostic the evidence was, called *misperception*, or failure to combine those diagnosticity assessments according to Bayes' rule, *misaggregation*. Aside from its theoretical interest, this dispute had considerable practical importance. If people can assess component probabilities but cannot combine them, then person–machine systems may be able to relieve them of the mechanical computations. Moreover, the system could incorporate an elicitation scheme that kept users from ever forgetting any component probabilities. However, if people are the only source of probabilities and they cannot assess them very well, then the machine may be just spinning its discs (Edwards, 1962).

Although the source of this conservatism was never determined, the hypotheses that were raised reflect the problems that could interfere with aggregation. Examples, details of which may be found in Slovic and Lichtenstein (1971) are (a) anchoring—people stay stuck to their previous estimates, (b) response bias—reluctance to give extreme responses pushes answers toward the center of the response range, (c) ceiling effect—fear of "using up" the probability scale makes people hedge their responses, (d) nonlinearity—although a given *D* should produce the same ratio of posterior odds to prior odds whenever it is received, people may try instead to make the differences between prior and posterior probability of *H* constant, and (e) response mode—probability assessments may be less optimal responses than the odds (or log odds) assessments for the same task; despite their formal equivalence, one response mode may be a more natural way for people to assess and express their knowledge about a particular problem.

In the end, this line of research was quietly abandoned without establishing the relative roles of these different factors. This cessation of activity seems to be partly due to the discovery of the base-rate fallacy, which represents the antithesis of conservatism and other phenomena that led researchers to conclusions such as the following: "It may not be unreasonable to assume that . . . the probability estimation task is too unfamiliar and complex to be meaningful" (Pitz, Downing, & Reinhold, 1967, p. 392). "Evidence to date seems to indicate that subjects are processing information in ways fundamentally different from Bayesian . . . models" (Slovic & Lichtenstein, 1971, p. 728). "In his evaluation of evidence, man is apparently not a conservative Bayesian; he is not Bayesian at all" (Kahneman & Tversky, 1972, p. 450).

Information Search

Obviously, Bayesian updating requires the collection of additional information beyond what was incorporated in the prior odds. Whether collection is contemplated at all should depend on whether one's a priori confidence in the truth of the hypothesis is adequate for deciding what to do and on the possibilities for additional data to change one's mind. When one would like to know more, the specific data collected should depend on the opportunities presented.

These opportunities can be conceptualized as questions that can elicit a set of possible answers, or D_i, each of which carries a message regarding the truth of the hypotheses. All other things being equal (e.g., the cost of asking), the most valuable questions are those that are expected to produce the most diagnostic answers. Conversely, one should never ask questions all of whose possible answers have likelihood ratios of 1. Such questions should change neither one's beliefs regarding the truth of the hypotheses nor one's choice of action based on those beliefs. Value-of-information analysis includes a variety of schemes for deciding how much one should spend for information in general and how to devise the most efficient sampling strategies. It considers such factors as the cost of asking, the consequences of the possible decisions, the a priori probability of receiving different possible answers, and the likelihood ratios associated with those answers (Brown, Kahr, & Peterson, 1974; Raiffa, 1968).

Why might someone disregard these considerations and ask questions whose answers cannot be diagnostic? Tradition is one possible reason. A datum may always have been collected, without serious analysis of what has been learned from it. Official forms and graduate school applications might be two familiar loci for nondiagnostic questions. These traditions may spawn or be spawned by beliefs about the kinds of evidence that are inherently more valuable. In various circles, secret, quantitative, or introspective information might have this special status of always meriting inquiry (Fischer, 1970). Misdirected search may occur also when people's task changes and they fail to realize that the questions that helped to evaluate the old hypotheses are no longer as effective in evaluating the new ones. For example, a psychiatric social worker who moved from private practice to a large public agency might require a whole new set of question-asking skills.

The factors leading to pointless questions can, in less extreme form, lead to inefficient ones—questions that provide some information, yet less than the maximum possible. These search problems may be aggravated by difficulties with other aspects of the inferential process. Value-of-information analysis requires a *preposterior* analysis, in which one anticipates what one will believe and what one will do if various possible data are observed. If people have difficulty assessing likelihood ratios for actual data, then they are unlikely to assess them properly for hypothetical data. A further obstacle to appraising the expected value of possible questions is the very hypotheticality of the question. When seen "in the flesh," information may have more or less impact than it was expected to have, in which case it is not clear whether to trust the actual or the anticipated impact (Fischhoff, Slovic, & Lichtenstein, 1979).

A noteworthy form of efficiency is ignoring the opportunity to ask potentially falsifying questions, ones whose answers have a reasonable chance of effectively ruling out some hypothesis (e.g., a physician failing to order a test that could eliminate the possibility of a particular disease). What constitutes a "reasonable chance" depends on the usual value-of-information analysis factors (e.g., the prior probability of the disease, the importance of its detection, the cost of the test).

An obvious danger in information search is inadvertently selecting an unrepresentative set of evidence and arriving at erroneous beliefs. When a proper sampling frame is available (e.g., for the Census), one can describe a variety of specific violations of representative sampling (Kish, 1965). If those biases can be characterized, then it is possible, in principle, to correct for them (Bar-Hillel, 1979). Such situations may, however, be relatively rare with Bayesian inference, where information can come from a variety of sources and in a variety of forms. Good sense then becomes the only guide to drawing and interpreting samples. As the history of scientific progress shows, it may take a fortuitous, if unpleasant, surprise to reveal an unintended bias in sampled information.[10]

Action

Bayesian inference is embedded in statistical decision theory. Its output, posterior odds, summarizes beliefs in a way that facilitates selecting the optimal course of action. Similarly, knowledge of the possible actions and their associated consequences is essential in determining what information to gather. Two Bayesian judges who contemplated different actions, or evaluated their consequences differently, might justifiably formulate different hypotheses and collect different data even though they agreed on the interpretation of all possible data.

Many of the difficulties that frustrate attempts to take prudent action on the basis of available knowledge are not unique to Bayesian inference. These include not having well-articulated values (i.e., not knowing what one wants), failing to think through all the consequences of different actions, and allowing one's preferences to be manipulated by the way in which problems are presented (Fischhoff, Slovic, & Lichtenstein, 1980; Rokeach, 1973; Tversky & Kahneman, 1981).

With other familiar problems, the Bayesian framework may offer an illuminating nomenclature and even some assistance. For example, people may forget that rejecting one option always means accepting another (if only the inaction option) that may prove even less attractive if it is examined. Conversely, accepting any option can mean forgoing others because there are not enough resources to go around. When evaluating an option, one must consider the *opportunity costs* of doing without the net benefits that would be gained by adopting other options (Vaupel & Graham, 1981). The Bayesian framework forces one to consider at least two options.

As mentioned earlier, the critical ratio provides a threshold for translating posterior odds into action: If they are above that threshold, act as though H were true; if they are below it, act as though $\bar{H}$ were true. The ratio is set by considering the consequences of being right and of being wrong in either case. Thus, it is the Bayesian way to relate uncertain knowledge to concrete actions by showing which action seems to be in one's best interest. In evaluating these actions, it is important to remember that (a) in the presence of uncertainty, the best action may not lead to the best outcome, (b) people must act on the basis of what they themselves believe at the time of decision, not what others believe or subsequently learn (Fischhoff, 1975), and (c) when uncertainty is great, there may be little difference in the apparent attractiveness of competing actions (von Winterfeldt & Edwards, 1982).

One danger in embedding inference in a decision-making framework is that it may encourage people to confuse "acting as though H is true" with "believing that H is true." Decision makers who confuse the two may not attend to signals indicating that their best guess about H was wrong and requires revision (Gärdenfors & Sahlin, 1982). Scientists who confuse the two may forget the uncertainties that they themselves acknowledged before offering a best-guess interpretation of experimental results.

Complications

In this presentation, the interpretation of a datum is complete once one has compared the two conditional probabilities comprising the likelihood ratio. Such appraisal assumes that the datum is taken at face value. More

[10] Deliberately unrepresentative sampling seems to be another possible bias to be included in this section. However, we argue in the discussion of Snyder and Swann (1978) that such biases are better conceptualized as problems of data interpretation than as problems of sampling.

sophisticated Bayesian models are available for situations in which that assumption seems dubious and the interpretation of data depends upon contextual factors. Two such elaborations deal with *source credibility* and *conditional independence*. These complications also point to the need for care in making claims of biased, or non-Bayesian, inferences.

Source Credibility

Every datum comes from some source. Knowing that source may, in principle, have quite diverse effects on the datum's interpretation. In common parlance one speaks about sources that have unusual or limited credibility as well as those that may attempt to mislead or may have been misled themselves. On the basis of detailed modeling of the informational properties of evidentiary situations that may arise in the courtroom, Schum (1980, 1981) has shown how source credibility information may reduce, enhance, or even reverse the diagnostic impact of a particular datum. The subtlety of Schum's models suggests both the difficulty of applying Bayesian inference properly and the pitfalls awaiting those trying to rely on intuition. Without understanding the impact of source credibility information, it is hard to know how it is or how it should be interpreted.

A common task in judgment research requires participants to decide whether a target individual belongs to Category A or Category B on the basis of a brief description and some base-rate information. These descriptions vary along dimensions such as the internal consistency of the information they contain and the ratio of relevant to irrelevant information. Typically, investigators have considered only the informational content of these messages when analyzing the impact that these variations should have and do have on behavior. In principle, however, each shift in content could signal a different level of credibility. If subjects are sensitive to these signals, they may choose to respond in ways that are at odds with those dictated by the informational content—and be justified in doing so.

For example, Manis, Dovalina, Avis, and Cardoze (1980), as well as Ginosar and Trope (1980), varied the consistency of the information in a description. With consistent profiles, all information that pointed toward any category pointed toward the same category; with inconsistent profiles, such diagnostic bits pointed in both directions. Subjects relied less on inconsistent information. This might reflect sensitivity to its apparently lower diagnosticity, or it might reflect doubt about its overall credibility. If one doubts that inconsistent people exist (Cooper, 1981), one may discount sources that have produced descriptions showing inconsistency. Both responses could be justified normatively and would lead to similar judgments. However, they suggest different psychological processes. The latter interpretation would mean that this is a situation in which people do understand the need to regress predictions based on unreliable information (Kahneman & Tversky, 1973).

In situations where the content of a message provides a cue regarding its validity, failure to consider that message may lead investigators to mistake sensitivity for bias. For example, in an experimental study of manuscript reviewing, Mahoney (1977) berated his scientist subjects for being more hospitable toward a fictitious study when its reported result confirmed the dominant hypothesis in their field than when it disconfirmed it. This differential receptiveness could reflect the stodginess and prejudices of normal science (Kuhn, 1962), which refuses to relinquish its pet beliefs. However, it could also reflect a belief that investigators who report disconfirming results tend to use inferior research methods (e.g., small samples leading to more spurious results), to commit common mistakes in experimental design, or, simply, to be charlatans. Mahoney himself might set a double standard if he were told that a study did or did not confirm the existence of telekinesis.

These reinterpretations are, of course, entirely speculative. To discipline them by fact, one needs to discover (a) how subjects structure the problem (e.g., their worries about source credibility) and (b) how they appraise the different components of the inferential model that they are using. For either describing or evaluating behavior, one must establish both what people believe and what they try to do with those beliefs (see also Wetzel, 1982).

Conditional Independence

The most general way of thinking about contextual effects is as an interaction between the meaning of two or more data. That is, one datum, D_i, creates a context that affects the interpretation of another datum, D_j. In Bayesian terms, such interactions are said to reflect *conditional nonindependence* because the conditional probability $P(D_j/H)$ is not necessarily equal to $P(D_j/H,D_i)$. As a result, one cannot compute the cumulative impact of a set of data simply by multiplying their respective likelihood ratios.

Source-credibility problems may be viewed as a special case of conditional nonindependence: Information about the source affects interpretation of the message. Conversely, the message may affect one's view of the credibility of the source. Conditional nonindependence is also the grounds for configural judgment, the focus of many studies of clinical diagnosis (Goldberg, 1968; Slovic & Lichtenstein, 1971). For the configural judge, the meaning of a particular cue depends upon the status of others (e.g., "That tone of voice suggests 'not suicidal' to me unless I know that it was spoken at midday"). The research record shows that, although clinicians claim to interpret cues configurally, firm evidence of configural judgment is hard to find. This discrepancy may reflect the insensitivity of the research tool, the inaccuracy of the clinicians' introspections about their own judgmental processes, or their failure to use their configural strategies consistently (Dawes, 1979; Slovic & Lichtenstein, 1971).

In studies of clinical diagnosis, both the relationships between cues and people's judgments of those relationships are modeled, typically by linear regression equations. Configural relations are represented in those equations by interaction terms. In the Bayesian model, conditional nonindependence is treated by assessing joint conditional probabilities that consider interrelated data simultaneously. Considering the subtleties of Schum's analyses of source-credibility problems, it is very difficult to model these relationships either in the world or in people's judgments. Indeed, this very complexity may mean that sets of interrelated data defy explicit normative modeling, leaving them the province of judgment (Navon, 1981).

For the expert analyst of evidence, this leads to the frustrating situation of having to take a best guess at what the data mean knowing that their mutual implications have not been understood. Because such complications are common, those frustrations will also be common. Although lay judges may face conditional nonindependence equally often, they may not always take account of it. Perhaps it is the norm to take evidence at face value unless some alarm is sounded by the evidence itself or by the source presenting it. If that is the case, then simple Bayesian models may prove as effective for hypothesis evaluation behavior as simple regression models proved for clinical judgment. As with the regression models, these Bayesian models could be interpreted literally or as "paramorphic models" capturing stimulus–response relationships without claiming to describe the underlying cognitive processes (Hoffman, 1960).

Limits

Just as there are practical limits to the informational complexities that can be treated adequately within the Bayesian framework, so there are value considerations that are best recognized and left alone. People are not always just acquiring knowledge for the sake of optimizing their actions. Some pursuits that are not sensibly accommodated in the Bayesian framework can lead to deliberately non-Bayesian behavior. For example, people may deliberately act suboptimally in the short run when they are pursuing long-run goals such as "maintaining social relations (e.g., preserving and cultivating information sources), gaining and sustaining recognition (e.g., exuding confidence where accountability is low), and being accepted (e.g., by passing up smart solutions that make one appear out of step)" (Fischhoff, 1981, p. 902). Thus, people may ask nondiagnostic questions in order to keep the conversation going, and they may pass up diagnostic ones because asking them seems untoward.

With sufficient ingenuity, such behavior could probably be translated into Bayesian terms so as to show that it is not just purposeful, but also optimal—in the sense of having the highest expected value of any course of action. Thus, for example, asking a question could be treated as an act that has consequences other than the cost of asking it and the information that it produces. These consequences might include the possible penalty of being censured for impertinence or the chance of producing a completely unexpected result leading to the creation of new hypotheses. Practically speaking, anticipating and analyzing such considerations would be very difficult. Theoretically speaking, the attempt to do so might have a very ad hoc character, as though the investigator were groping for causes that might conceivably shape people's hypothesis evaluation. To be useful, these interpretations must walk a tightrope between giving people too little credit and giving them too much. At the former extreme, any behavior for which the investigator finds no ready Bayesian expression reflects cognitive incompetence. At the other extreme, people do whatever is right for them, and the observer's task is to determine what it is that they have managed to optimize (Cohen [and commentary], 1981; Fischhoff, Goitein, & Shapira, 1982; Hogarth, 1981).

Implications and Reinterpretations

Prescriptively, the Bayesian approach provides a general model of how people should make sequential inferences. Using the model descriptively requires a choice between two strategies. One strategy is to assume that people are intuitive Bayesians and explore how they use the model. Alternatively, one can assume that they do not use the model but that their judgments can be described in terms of systematic departures from it (as when we characterized a number of phenomena as special cases of failing to consider the denominator in the likelihood ratio). In the course of explicating the Bayesian model, we have used both strategies, at times applying them to the same observed behavior. In some of these cases, that behavior could be viewed either as Bayesian or non-Bayesian, depending upon what one assumes about what people believe and what problem they are solving.

We use this framework now to analyze the behavior that has been observed in a number of additional studies. Typically, it casts a somewhat different light on what the subjects in those studies were doing and should have been doing than was cast by the original authors. In some cases, this reinterpretation shows commonalities in effects that had appeared to be distinct. In others, it reveals the differences in tasks that had gone under the same label. In particular, it shows how the term *confirmatory bias* has been applied to a variety of phenomena that may be described more succinctly in terms of the Bayesian model.

Nisbett, Zukier, and Lemley (1981)

The stimuli in Nisbett et al. were thumbnail descriptions of fictional individuals. The hypotheses were of the form "the individual belongs to category A" (e.g., is a child abuser). Stimuli and hypotheses were designed to test the authors' integration of Kahneman and Tversky's (1972) work on representativeness with Tversky's (1977) work on similarity judgments. According to the former, the judged probability of an individual belonging to a category depends on the judged similarity between the salient features of the individual's description and of the category stereotype. According to the latter, judged similarity should increase with the number of salient features common to both the individual's description and the category stereotype; it should decrease with the number of features unique to each.

Nisbett et al.'s descriptions consisted of one or two diagnostic features, each pointing toward one of the two possible prediction categories, and a varying number of nondiagnostic features, pointing toward neither category. The authors found that as the number of nondiagnostic features increased, the impact of the diagnostic features decreased. Thus, the nondiagnostic data dilute the effect of the diagnostic data, leading to less confident predictions. As the authors note, this dilution could only be justified normatively if subjects perceived, what would be called here, conditional nonindependence among the descriptors; that is, if the nondiagnostic information somehow mitigated the impact of the diagnostic information, for example, by reminding subjects how complex people are and hence how unreasonable it is to make confident predictions on the basis of a single feature.

Although the possible normative interpretations of these results are quite clear, the role of similarity judgments in them is not. As described by Tversky (1977), similarity judgments concern a match between an individual and a category. If subjects rely on representativeness, then the judged probability of an individual belonging to any category depends on the apparent individual–category similarity. Nisbett et al. extend these notions to cover the case of two possible categories by assuming that people form a subjective likelihood ratio whose numerator and denominator are derived separately by representativeness from judged individual–category similarity.

Nisbett et al. designed their stimuli so that nondiagnostic data belonged to neither category stereotype. As a result, adding a nondiagnostic datum to a diagnostic datum should produce a description that is less similar to both category stereotypes than the single-datum description. The judged probability of getting such a description given that the individual belongs to each category should be correspondingly lower. If both conditional probabilities are reduced by equivalent amounts, then there should be no reduction in diagnosticity, hence no dilution effect. Given that the addition of nondiagnostic information did reduce the extremity of judgments, one must forfeit some component of the above account. One solution is to drop the assumption that subjects analyze evidence by forming subjective likelihood ratios. Rather, here as elsewhere, subjects ignore the denominator when considering diagnosticity. By doing so, they would not notice that the nondiagnostic information also reduced the match between the description and the competing category.

Such neglect of the denominator would have two additional implications for the interpretation of Nisbett et al.'s study. One is that pretest subjects would not be judging diagnosticity in the conventional sense when they appraised "how helpful the information would be for prediction" (p. 255). The second is that encouraging people to rely on representativeness in the manner suggested by Nisbett et al. might actually improve their inference. Asking whether *D* is more representative of A or B would mean making a judgment that has at least the elements of the likelihood ratio—a datum and two competing hypotheses—if not necessarily in the proper relationship to each other. People might also be cautioned about a problem that arises with reliance on this strategy, the tendency to neglect the priors whenever one can detect any degree of differential representativeness (Bar-Hillel & Fischhoff, 1981; Ginosar & Trope, 1980).

Snyder and Swann (1978)

In a series of studies, Snyder and Swann (1978) had subjects select questions that "would provide them with the information to best test the hypothesis" (p. 1204) that a target individual whom they were about to interview was an introvert or an extravert. In the subjects' choice of questions, the authors claimed to have demonstrated a new bias, namely "an erroneous tendency to search for evidence that would tend to confirm the hypothesis under scrutiny" (p. 1203).

From a Bayesian perspective, however, it is unclear how the choice of question could be biased toward the confirmation of a hypothesis. Mathematically, it is not possible to ask questions all of whose possible answers support a particular hypothesis. Thus, there is no confirmatory bias in the sense of selecting data that inevitably favor a particular hypothesis. Snyder and Swann apparently felt that their subjects asked questions whose answers they could anticipate and would interpret as supporting the focal hypothesis. If the answer, D_i, to a question is predictable, then $P(D_i) = 1 = P(D_i/H) = P(D_i/\bar{H})$. That is, the question is nondiagnostic and the search process is inefficient. If D_i is taken as evidence supporting H, then the problem lies with the interpretation, which may reflect neglect of the denominator in the likelihood ratio.

A less extreme form of this exploitation of predictability might be seen in a scientist who repeatedly replicates the same experiment, which typically produces the same observation, D, which is more likely given H than given $\bar{H}$. Each observation of D would not, however, provide an independent and equal contribution to affirming H. An observation is informative only to the extent that it is unpredictable, hence capable of affirming or disaffirming the hypothesis. With each replication of the experiment, $P(D)$ increases. There is a corresponding decrease in the disparity between $P(D/H)$ and $P(D/\bar{H})$ that is necessary for a diagnostic likelihood ratio. After many replications, $P(D)$ and the likelihood ratio both approach 1, making future experimentation uninformative. An analogous way to look at this problem is to view the likelihood ratio as a measure of association defined on a 2 × 2 table whose rows and columns are $(D, \bar{D})$ and $(H, \bar{H})$, respectively. As the marginals of one of the variables (here, D) become more extreme, the value of the measure of association tends to decrease.

Perhaps the only way to bias the search toward accepting the focal hypothesis, H, would be to ask questions for which both $P(D/H)$ and $P(D/\bar{H})$ were expected to be high, knowing that one would subsequently ignore $P(D/\bar{H})$. Such a biased search in-the-service-of biased interpretation could, of course, only be exhibited successfully when $P(D)$ is high (i.e., D is likely to be observed whether or not H is true). If subjects tried it when $P(D)$ was low, then it would represent a disconfirmation bias: Subjects would look only at $P(D/H)$, which would be low, thereby reducing their confidence in H.

In Snyder and Swann's experiment, subjects asked to test the hypothesis that the target individual was an extravert were counted as biased if they chose "confirmatory" questions such as "What would you do if you wanted to liven up a party?" rather than "disconfirmatory" questions such as "What factors make it hard for you to really open up to people?" or "neutral" questions, such as, "What are your career goals?" The nonneutral questions have two noteworthy properties. One is that they are nondiagnostic insofar as they would elicit similar responses from both introverts and extraverts (and unless there are particularly introverted ways to liven up parties). Although such questions are inefficient, their selection constitutes no bias.toward confirmation. As there were only 5 neutral questions in the set of 26 possibilities, subjects still had to choose many nonneutral ones in their set of 12 test questions. A second peculiar property of the nonneutral questions is that they are phrased in a conditional way that assumes the category membership of the recipient. For example, it would seem awkward or untoward to ask an introvert "What would you do if you wanted to liven up a party?" (Trope & Basok, 1982). Perhaps the only way to get any useful information with such a question would be if respondents rejected its premises where appropriate (e.g., "What do you mean? I never want to liven up parties!").[11] Subjects' preference among nonneutral questions for those relating to the focal hypothesis ("confirmatory" questions in Snyder and Swann's terminology) might be traced to a desire for compatibility with that hypothesis, or it could be that subjects treat the focal hypothesis as true and ask questions that would be the least awkward socially.

Wason (1960, 1968)

Card task. Some analogous phenomena may be seen in studies of how people test what might be called *formal* or *logical* hypotheses (Evans, 1982; Wason, 1960, 1968). These are categorical statements, such as "All *A*s are *B*s," that may be disconfirmed by a single counterexample (e.g., an A that is not a B). In one popular experiment, subjects are told that each of four cards has either A or $\bar{A}$ on one side and either B or $\bar{B}$ on the other

[11] Only in Experiment 2 did Snyder and Swann's subjects actually conduct the promised interview with the target person. Other subjects who listened to tapes of these interviews judged respondents to predominantly extravert questions to be more extraverted than respondents to predominantly introvert questions. One might concur with Snyder and Swann's speculation that the biased sample of questions evoked a biased sample of reported behavior, or one might speculate that the form of the interviewees' responses reflected the conditioning presumptions of the questions they were asked. That is, it may be relatively easy to tell when people are answering questions that "would typically be asked of people *already known* to be extraverts" (Snyder & Swann, 1978, p. 1204; emphasis in original).

side. They are then shown the four cards, whose exposed sides show, respectively, A, $\bar{A}$, B, and $\bar{B}$. Subjects' task is to choose the cards they would turn over in order to test the hypothesis that "All As are Bs." Even though turning over either the A or the $\bar{B}$ could falsify the hypothesis, most people "waste" one of their choices on the B card, which is necessarily inconclusive. From our perspective, this bias is an example both of asking a nondiagnostic question (B) and of failing to ask a potentially falsifying question ($\bar{B}$).

Such a task would constitute a special case of Bayesian inference in at least two senses. One is that with logical hypotheses it is possible to observe data that unambiguously falsify or verify an hypothesis (i.e., that have a likelihood ratio of 0 or infinity). With empirical hypotheses, regarding real-life events, such certitude is rare or impossible. Indeed, one might speculate that subjects fail to ask potentially falsifying questions because they are unaccustomed to testing logical hypotheses and to having falsification possible. Second, the artificiality of the problem gives subjects no basis for assigning most of the probabilities involved. Many different prior odds could be defended by imputing a likelihood that the experimenter would focus attention on a true hypothesis. Except for data that would disconfirm H or $\bar{H}$, it is hard to defend any value of the likelihood ratio, insofar as any value other than 0 and infinity requires some arbitrary assumptions about the correlation between A-ness and B-ness in the artificial universe that the experimenter created.

Triads task. In another task devised by Wason (1968), subjects are told that the numbers 2, 4, and 6 conform to a simple rule that the experimenter has in mind; their task is to guess that rule. As an aid, they may ask whether additional number triads of their own choosing conform to the rule. In such a logical task, the validity of a hypothesized rule cannot be proven. Even if a rule fits all conforming triads, there is no guarantee that some other rule might not also fit or that some future triad might not violate it. It is only subjects' suppositions about the sorts of rules that the experimenter might use and the negligible penalty for guessing wrong that would enable strictly Bayesian subjects to stop gathering information, that is, stop proposing triads and guess that a particular hypothesis is correct.

The experimenter's unannounced rule is "numbers increasing in value." The first rule that comes to most people's minds is apparently "sequential even numbers." The modal first triad is something like 8, 10, 12. Called a sign of "verification bias" by Wason, this response pattern may be explained by several different accounts.

The first account traces this apparent search bias to an interpretation bias. It argues that subjects believe that being told that 8, 10, 12 fits the rule will prove that "sequential even numbers" is the correct hypothesis, whereas being told that it does not fit will prove that "sequential even numbers" is not the correct hypothesis. Hence, they propose 8, 10, 12, expecting to get a definitive answer either way. This flaw in logical inference could be given a special name, such as *mistaking affirmation for confirmation.* However, it is most parsimoniously regarded as yet another instance of ignoring alternative hypotheses when evaluating evidence. For the datum "yes, it conforms" and the hypothesis "the rule is sequential even numbers," $P(D/H) = 1$. Hence, that answer cannot reduce one's confidence in the truth of H. The evidence becomes, of course, less conclusive when one realizes that many alternative hypotheses, including the experimenter's own, would evoke the same response and have the same associated conditional probability.

A second account holds that subjects understand the meaning of disconfirming evidence and that they view their task as comparing their favored hypothesis with its complement, all other rules. The question they ask (8, 10, 12) does, in fact, produce information that is diagnostic for this comparison, that is, its answer should change their confidence in their hypothesis. If they can be faulted, it is for inefficiency, asking a question whose expected answer will tell them very little.

In the third account, subjects compare their favored hypothesis with another specific hypothesis, acting for the moment as though those two exhausted the universe of possibilities. Although they might be faulted for making such a false assumption, subjects might still consider this simplification a useful fiction. A more "proper" strategy is to compare each hypothesis with its complement and then, should falsifying evidence be found, start afresh with a new hypothesis. However, in a situation where any one of an infinite number of hypotheses could be the true one, people may prefer to compare pairs of hypotheses sequentially; at each round, the less likely hypothesis is discarded and the more likely one is then compared with the next contender. Such a comparative strategy would be similar in spirit to *sophisticated falsificationism* (Lakatos, 1970), where one holds onto the hypothesis that seems most correct until a better candidate comes along, even to the point of retaining a hypothesis for which inconsistent data are known to exist. The "proper" strategy resembles *naive falsificationism* (Popper, 1972), where one focuses on a single hypothesis, doing everything possible to disprove it without regard for what might take its place.

Subjects using this comparative strategy might still be faulted for poor triad selection. For any pair of hypotheses, they should ask about triads for which an affirmative answer will falsify one hypothesis and a negative answer will falsify the other. Thus, 8, 10, 12 is a poor triad if the two hypotheses are "sequential even numbers" and "numbers increasing in value." The experimenter's response would be "yes" if either were the correct hypothesis. On the other hand, it would be a fine choice if the two hypotheses were "sequential even numbers" and "numbers less than 7." Without eliciting subjects' hypotheses, it is hard to tell whether to fault them for using the comparative strategy or for using it inefficiently by choosing nondiagnostic triads.

Conclusion

Bayesian inference was originally developed as a prescriptive model. Its advocates believe that when one evaluates hypotheses it is useful to identify each element in the Bayesian model with the corresponding elements in one's thinking. Such identification ensures that all necessary elements have been considered and that they have been put in the proper relationship to one another. Although the model can help people to structure their thinking, by itself it cannot show how to formulate hypotheses, assess component probabilities, or set the critical ratio. Performing these operations requires a substantive understanding of the problem at hand.

Like other prescriptive models, the Bayesian scheme assumes that people could follow its dictates if they tried and if they were given some feasible level of assistance. However, it makes no statement regarding how people actually do make decisions. The descriptive potential of the Bayesian model lies in the framework that it provides for the primitives and processes of people's intuitive inference. This article attempts to use this potential by identifying the kinds of systematic deviations from the Bayesian model that could, in principle, be observed. Illustrative examples of most of these biases are found in the research literature, some collected within a Bayesian framework, some not.

Once developed, this descriptive model was applied to interpret a variety of existing studies. Although the results of this analysis are specific to those studies, a number of general conclusions emerge:

1. In some situations, apparently diverse effects prove to be special cases of a single judgmental bias. The most powerful of these "metabiases" is the tendency to ignore $P(D/\bar{H})$ when evaluating evidence.

2. In some situations, a variety of different phenomena have been confused under a common title. *Confirmation bias,* in particular, has proven to be a catch-all phrase incorporating biases in both information search and interpretation. Because of its excess and conflicting meanings, the term might best be retired.

3. In all situations, a careful appraisal of how judges have interpreted the tasks posed to them is needed before making any assertions regarding how, if at all, their judgment is biased. When speculating about how people may have construed a task, it is important to strike a balance between exaggerating the extent to which we or they are all-knowing. Neither our understanding of people nor our

ability to help them is served by uncritically assuming either that there is no way for them to justify behavior that seems suboptimal to us or that there is a hidden method to any apparent madness that they exhibit.

Reference Notes

1. U.S. Nuclear Regulatory Commission. *Fault tree handbook* (NUREG-0492). Washington, D.C.: Author, 1981.
2. U.S. Nuclear Regulatory Commission. *Reactor safety study: An assessment of accident risks in U.S. commercial nuclear power plants* (WASH-1400, NUREG-75/014). Washington, D.C.: Author, 1975.
3. U.S. Nuclear Regulatory Commission. *Risk assessment review group report to the U.S. Nuclear Regulatory Commission* (NUREG/CR-0400). Washington, D.C.: Author, 1978.

References

Bar-Hillel, M. The role of sample size in sample evaluation. *Organizational Behavior and Human Performance,* 1979, *24,* 245–257.

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

Bar-Hillel, M., & Fischhoff, B. When do base rates affect predictions? *Journal of Personality and Social Psychology,* 1981, *41,* 671–680.

Beyth-Marom, R. How probable is probable? Numerical translation of verbal probability expressions. *Journal of Forecasting,* 1982, *1,* 257–269.

Borgida, E., & Brekke, N. The basc-rate fallacy in attribution and prediction. In J. H. Harvey, W. J. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 3). Hillsdale, N.J.: Erlbaum, 1981.

Brown, R. V., Kahr, A. S., & Peterson, C. *Decision analysis for the manager.* New York: Holt, Rinehart & Winston, 1974.

Cohen, J. Can human irrationality be experimentally demonstrated? *The Behavioral and Brain Sciences,* 1981, *4,* 317–370.

Cooper, W. H. Ubiquitous halo. *Psychological Bulletin,* 1981, *90,* 218–244.

Dawes, R. M. The robust beauty of improper linear models in decision making. *American Psychologist,* 1979, *34,* 571–582.

DeFinetti, B. Probability: Beware of falsifications! *Scientia,* 1976, *3,* 283–303.

Diaconis, P., & Zabell, S. L. Updating subjective probability. *Journal of the American Statistical Association,* 1982, *77,* 822–829.

Doherty, M. E., Mynatt, C. R., Tweney, R. D., & Schiavo, M. D. Pseudodiagnosticity. *Acta Psychologica,* 1979, *43,* 111–121.

Doyle, A. C. *The memoirs of Sherlock Holmes.* London: Murray & Cape, 1974. (Originally published 1893.)

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

Edwards, W. Dynamic decision theory and probabilistic information processing. *Human Factors,* 1962, *4,* 59–73.

Edwards, W. Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968.

Edwards, W., Lindman, H., & Savage, L. J. Bayesian statistical inference for psychological research. *Psychological Review,* 1963, *70,* 193–242.

Einhorn, H. J., & Hogarth, R. M. Prediction, diagnosis and causal thinking in forecasting. *Journal of Forecasting,* 1982, *1,* 23–36.

Evans, J. St. B. T. *The psychology of deductive reasoning.* London: Routledge & Kegan Paul, 1982.

Fischer, D. H. *Historians' fallacies.* New York: Harper & Row, 1970.

Fischhoff, B. Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance,* 1975, *1,* 288–299.

Fischhoff, B. Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance,* 1977, *3,* 349–358.

Fischhoff, B. For those condemned to study the past: Reflections on historical judgment. In R. A. Shweder & D. W. Fiske (Eds.), *New directions for methodology of behavior science: Fallible judgment in behavioral research.* San Francisco: Jossey-Bass, 1980.

Fischhoff, B. Inferential interference (Review of *Human inference: Strategies and shortcomings of social judgment,* by R. Nisbett & L. Ross). *Contemporary Psychology,* 1981, *26,* 901–903.

Fischhoff, B. Debiasing. In D. Kahneman, P. Slovic, & A. Tversky, (Eds.), *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press, 1982.

Fischhoff, B., Goitein, B., & Shapira, Z. The experienced utility of expected utility approaches. In N. Feather (Ed.), *Expectancy, incentive and action.* Hillsdale, N.J.: Erlbaum, 1982.

Fischhoff, B., Slovic, P., & Lichtenstein, S. Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance,* 1978, *4,* 330–344.

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

Fischhoff, B., Slovic, P., & Lichtenstein, S. Knowing what you want: Measuring labile values. In T. Wallsten (Ed.), *Cognitive processes in choice and decision behavior.* Hillsdale, N.J.: Erlbaum, 1980.

Forer, B. The fallacy of personal validation: A classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology,* 1949, *44,* 118–123.

Gärdenfors, P., & Sahlin, N.-E. Unreliable probabilities, risk taking, and decision making. *Synthese,* 1982, *53,* 361–386.

Ginosar, Z., & Trope, Y. The effects of base rates and individuating information on judgments about another person. *Journal of Experimental Social Psychology,* 1980, *16,* 228-242.

Goldberg, L. R. Simple models or simple processes? Some research in clinical judgment. *American Psychologist,* 1968, *23,* 486-496.

Good, I. J. *Probability and the weighting of evidence.* New York: Hafner, 1950.

Green, A. E., & Bourne, A. J. *Reliability technology.* New York: Wiley Interscience, 1972.

Hoffman, P. J. The paramorphic representation of clinical judgment. *Psychological Bulletin,* 1960, *47,* 116-131.

Hogarth, R. M. Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin,* 1981, *90,* 191-217.

How indexation builds in inflation. *Business Week,* November 12, 1979, pp. 114-116.

Hyman, R. Cold reading. *Zetetic* (The Skeptical Inquirer), 1977, *1*(2), 18-37.

Jeffrey, R. *The logic of decision.* New York: McGraw-Hill, 1965.

Jeffrey, R. Probable knowledge. In I. Lakatos (Ed.), *The problem of inductive logic.* Amsterdam: North Holland, 1968.

Kahneman, D., & Tversky, A. Subjective probability: A judgment of representativeness. *Cognitive Psychology,* 1972, *3,* 430-454.

Kahneman, D., & Tversky, A. On the psychology of prediction. *Psychological Review,* 1973, *80,* 237-251.

Kahneman, D., & Tversky, A. Intuitive predictions: Biases and corrective procedures. *TIMS Studies in Management Science,* 1979, *12,* 313-327.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.), *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press, 1982.

Kish, L. *Survey sampling.* New York: Wiley, 1965.

Koriat, A., Lichtenstein, S., & Fischhoff, B. Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory,* 1980, *6,* 107-118.

Kuhn, T. *Structure of scientific revolutions.* Princeton, N.J.: Princeton University Press, 1962.

Kyburg, H. E., & Smokler, H. E. (Eds.), *Studies in subjective probability.* Huntington, New York: Krieger, 1980.

Lakatos, I. Falsification and scientific research programs. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of scientific knowledge.* New York: Cambridge University Press, 1970.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press, 1982.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory,* 1978, *4,* 551-578.

Lindley, D. V. *Introduction to probability and statistics from a Bayesian viewpoint.* Cambridge, England: Cambridge University Press, 1965.

Lindley, D. V., Tversky, A., & Brown, R. V. On the reconciliation of probability assessments. *Journal of the Royal Statistical Society* (Series A), 1979, *142,* Pt. 2, 146-180.

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

Mahoney, M. J. Publication prejudices. *Cognitive Therapy and Research,* 1977, *1,* 161-175.

Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. Base rates can affect individual predictions. *Journal of Personality and Social Psychology,* 1980, *38,* 231-240.

Mehle, T., Gettys, C. V., Manning, C., Baca, S., & Fisher, S. The availability explanation of excessive plausibility assessments. *Acta Psychologica,* 1981, *49,* 127-140.

Moskowitz, H., Sarin, R. K. Improving the consistency of conditional probability assessments for long range forecasting and decision making. *Management Science,* in press.

Murphy, A. H. Scalar and vector partitions of the probability score: Two-stage situation. *Journal of Applied Meteorology,* 1972, *11,* 273-282.

Navon, D. Statistical and metastatistical considerations in analysing the desirability of human Bayesian conservatism. *British Journal of Mathematical & Statistical Psychology,* 1981, *34,* 205-212.

Nisbett, R., & Ross, L. *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, N.J.: Prentice-Hall, 1980.

Nisbett, R. E., Zukier, H., & Lemley, R. E. The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology,* 1981, *13,* 248-277.

Novick, M. R., & Jackson, P. E. *Statistical methods for educational and psychological research.* New York: McGraw-Hill, 1974.

O'Leary, M. K., Coplin, W. D., Shapiro, H. B., & Dean, D. The quest for relevance. *International Studies Quarterly,* 1974, *18,* 211-237.

Phillips, L. D. *Bayesian statistics for social scientists.* London: Nelson, 1973.

Pitz, G. F., Downing, L., & Reinhold, H. Sequential effects in the revision of subjective probabilities. *Canadian Journal of Psychology,* 1967, *21,* 381-393.

Popper, K. R. *Objective knowledge.* Oxford, England: Clarendon, 1972.

Raiffa, H. *Decision analysis.* Reading, Mass.: Addison-Wesley, 1968.

Reyna, V. F. The language of possibility and probability: Effects of negation on meaning. *Memory and Cognition,* 1981, *9,* 642-650.

Rokeach, M. *The nature of human values.* New York: Free Press, 1973.

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

Schlaifer, R. *Analysis of decisions under uncertainty.* New York: McGraw-Hill, 1969.

Schum, D. Current developments in research on cascaded inference processes. In T. Wallsten (Ed.), *Cognitive processes in choice and decision behavior.* Hillsdale, N.J.: Erlbaum, 1980.

Schum, D. Sorting out the effects of witness sensitivity and response criterion placement upon the inferential value of testimonial evidence. *Organizational Behavior and Human Performance,* 1981, *27,* 153-196.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance,* 1971, *6,* 649–744.

Snyder, M., & Swann, W. B. Hypothesis testing processes in social interaction. *Journal of Personality and Social Psychology,* 1978, *36,* 1202–1212.

Sue, S., Smith, R. E., & Caldwell, C. Effects of inadmissible evidence on the decisions of simulated jurors: A moral dilemma. *Journal of Applied Social Psychology,* 1973, *3,* 345–353.

Trope, Y., & Basok, M. Confirmatory and diagnosing strategies in social-information gathering. *Journal of Personality and Social Psychology,* 1982, *43,* 22–34.

Troutman, C. M., & Shanteau, J. Inferences based on nondiagnostic information. *Organizational Behavior and Human Performance,* 1977, *19,* 43–55.

Tversky, A. Features of similarity. *Psychological Review,* 1977, *84,* 327–352.

Tversky, A., & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science,* 1974, *185,* 1124–1131.

Tversky, A., & Kahneman, D. Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology.* Hillsdale, N.J.: Erlbaum, 1980.

Tversky, A., & Kahneman, D. The framing of decisions and the rationality of choice. *Science,* 1981, *211,* 453–458.

Vaupel, J., & Graham, J. Eggs in your bier. *The Public Interest,* 1981, *61,* 3–17.

Wason, P. C. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology,* 1960, *12,* 129–140.

Wason, P. C. Reasoning about a rule. *Quarterly Journal of Experimental Psychology,* 1968, *23,* 273–281.

Wetzel, C. G. Self-serving biases in attribution: A Bayesian analysis. *Journal of Personality & Social Psychology,* 1982, *43,* 197–209.

von Winterfeldt, D., & Edwards, W. Costs and payoffs in perceptual research. *Psychological Bulletin,* 1982, *91,* 609–622.

Received May 5, 1982
Revision received March 16, 1983 ■