Psychological Review
1987, Vol. 94, No. 2, 211-228

Copyright 1987 by the American Psychological Association, Inc.
0033-295X/87/$00.75

Confirmation, Disconfirmation, and Information in Hypothesis Testing

Joshua Klayman and Young-Won Ha
Center for Decision Research, Graduate School of Business, University of Chicago

Strategies for hypothesis testing in scientific investigation and everyday reasoning have interested both psychologists and philosophers. A number of these scholars stress the importance of disconfirmation in reasoning and suggest that people are instead prone to a general deleterious "confirmation bias." In particular, it is suggested that people tend to test those cases that have the best chance of verifying current beliefs rather than those that have the best chance of falsifying them. We show, however, that many phenomena labeled "confirmation bias" are better understood in terms of a general *positive test strategy.* With this strategy, there is a tendency to test cases that are expected (or known) to have the property of interest rather than those expected (or known) to lack that property. This strategy is not equivalent to confirmation bias in the first sense; we show that the positive test strategy can be a very good heuristic for determining the truth or falsity of a hypothesis under realistic conditions. It can, however, lead to systematic errors or inefficiencies. The appropriateness of human hypothesis-testing strategies and prescriptions about optimal strategies must be understood in terms of the interaction between the strategy and the task at hand.

A substantial proportion of the psychological literature on hypothesis testing has dealt with issues of confirmation and disconfirmation. Interest in this topic was spurred by the research findings of Wason (e.g., 1960, 1968) and by writings in the philosophy of science (e.g., Lakatos, 1970; Platt, 1964; Popper, 1959, 1972), which related hypothesis testing to the pursuit of scientific inquiry. Much of the work in this area, both empirical and theoretical, stresses the importance of disconfirmation in learning and reasoning. In contrast, human reasoning is often said to be prone to a "confirmation bias" that hinders effective learning. However, confirmation bias has meant different things to different investigators, as Fischhoff and Beyth-Marom point out in a recent review (1983). For example, researchers studying the perception of correlations have proposed that people are overly influenced by the co-occurrence of two events and insufficiently influenced by instances in which one event occurs without the other (e.g., Arkes & Harkness, 1983; Crocker, 1981; Jenkins & Ward, 1965; Nisbett & Ross, 1980; Schustack & Sternberg, 1981; Shaklee & Mims, 1982; Smedslund, 1963; Ward & Jenkins, 1965). Other researchers have suggested that people tend to discredit or reinterpret information counter to a hypothesis they hold (e.g., Lord, Ross, & Lepper, 1979; Nisbett & Ross, 1980; Ross & Lepper, 1980) or they may conduct biased tests that pose little risk of producing disconfirming results (e.g., Snyder, 1981; Snyder & Campbell, 1980; Snyder & Swann, 1978).

The investigation of hypothesis testing has been concerned with both descriptive and prescriptive issues. On the one hand, researchers have been interested in understanding the processes by which people form, test, and revise hypotheses in social judgment, logical reasoning, scientific investigation, and other domains. On the other hand, there has also been a strong implication that people are doing things the wrong way and that efforts should be made to correct or compensate for the failings of human hypothesis testing. This concern has been expressed with regard to everyday reasoning (e.g., see Bruner, 1951; Nisbett & Ross, 1980) as well as professional scientific endeavor (e.g., Mahoney, 1979; Platt, 1964).

In this article, we focus on hypotheses about the factors that predict, explain, or describe the occurrence of some event or property of interest. We mean this broadly, to include hypotheses about causation ("Cloud seeding increases rainfall"), categorization ("John is an extrovert"), prediction ("The major risk factors for schizophrenia are . . ."), and diagnosis ("The most diagnostic signs of malignancy are . . ."). We consider both descriptive and prescriptive issues concerning information gathering in hypothesis-testing tasks. We include under this rubric tasks that require the acquisition of evidence to determine whether or not a hypothesis is correct. The task may require the subject to determine the truth value of a given hypothesis (e.g., Jenkins & Ward, 1965; Snyder & Campbell, 1980; Wason, 1966), or to find the one true hypothesis among a set or universe of possibilities (e.g., Bruner, Goodnow, & Austin, 1956; Mynatt, Doherty, & Tweney, 1977, 1978; Wason, 1960, 1968).

The task known as rule discovery (Wason, 1960) serves as the basis for the development of our analyses, which we later extend to other kinds of hypothesis testing. We first examine what "confirmation" means in hypothesis testing. Different senses of confirmation have been poorly distinguished in the literature, contributing to misinterpretations of both empirical findings

This work was supported by Grant SES-8309586 from the Decision and Management Sciences program of the National Science Foundation. We thank Hillel Einhorn, Ward Edwards, Jackie Gnepp, William Goldstein, Steven Hoch, Robin Hogarth, George Loewenstein, Nancy Pennington, Jay Russo, Paul Schoemaker, William Swann, Tom Trabasso, Ryan Tweney, and three anonymous reviewers for invaluable comments on earlier drafts.

Correspondence concerning this article should be addressed to Joshua Klayman, Graduate School of Business, University of Chicago, 1101 East 58th Street, Chicago, Illinois 60637.

and theoretical prescriptions. We propose that many phenomena of human hypothesis testing can be understood in terms of a general *positive test strategy.* According to this strategy, you test a hypothesis by examining instances in which the property or event is expected to occur (to see if it does occur), or by examining instances in which it is known to have occurred (to see if the hypothesized conditions prevail). This basic strategy subsumes a number of strategies or tendencies that have been suggested for particular tasks, such as confirmation strategy, verification strategy, matching bias, and illicit conversion. As some of these names imply, this approach is not theoretically proper. We show, however, that the positive test strategy is actually a good all-purpose heuristic across a range of hypothesis-testing situations, including situations in which rules and feedback are probabilistic. Under commonly occurring conditions, this strategy can be well suited to the basic goal of determining whether or not a hypothesis is correct.

Next, we show how the positive test strategy provides an integrative frame for understanding behavior in a variety of seemingly disparate domains, including concept identification, logical reasoning, intuitive personality testing, learning from outcome feedback, and judgment of contingency or correlation. Our thesis is that when concrete, task-specific information is lacking, or cognitive demands are high, people rely on the positive test strategy as a general default heuristic. Like any all-purpose strategy, this may lead to a variety of problems when applied to particular situations, and many of the biases and errors described in the literature can be understood in this light. On the other hand, this general heuristic is often quite adequate, and people do seem to be capable of more sophisticated strategies when task conditions are favorable.

Finally, we discuss some ways in which our task analysis can be extended to a wider range of situations and how it can contribute to further investigation of hypothesis-testing processes.

Confirmation and Disconfirmation in Rule Discovery

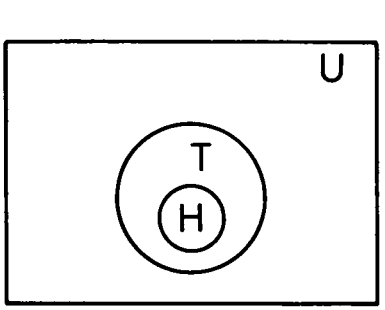
The Rule Discovery Task

Briefly, the rule discovery task can be described as follows: There is a class of objects with which you are concerned; some of the objects have a particular property of interest and others do not. The task of rule discovery is to determine the set of characteristics that differentiate those with this target property from those without it. The concept identification paradigm in learning studies is a familiar example of a laboratory rule-discovery task (e.g. Bruner, Goodnow, & Austin, 1956; Levine, 1966; Trabasso & Bower, 1968). Here, the objects may be, for example, visual stimuli in different shapes, colors, and locations. Some choices of stimuli are reinforced, others are not. The learner's goal is to discover the rule or "concept" (e.g., red circles) that determines reinforcement.

Wason (1960) was the first to use this type of task to study people's understanding of the logic of confirmation and disconfirmation. He saw the rule-discovery task as representative of an important aspect of scientific reasoning (see also Mahoney, 1976, 1979; Mynatt et al., 1977, 1978; Simon, 1973). To illustrate the parallel between rule discovery and scientific investigation, consider the following hypothetical case. You are an astrophysicist, and you have a hypothesis about what kinds of stars develop planetary systems. This hypothesis might be derived from a larger theory of astrophysics or may have been induced from past observation. The hypothesis can be expressed as a rule, such that those stars that have the features specified in the rule are hypothesized to have planets and those not fitting the rule are hypothesized to have no planets. We will use the symbol R_H for the hypothesized rule, H for the set of instances that fit that hypothesis, and $\bar{H}$ for the set that do not fit it. There is a domain or "universe" to which the rule is meant to apply (e.g., all stars in our galaxy), and in that domain there is a target set (those stars that really do have planets). You would like to find the rule that exactly specifies which members of the domain are in the target set (the rule that describes exactly what type of stars have planets). We will use T for the target set, and R_T for the "correct" rule, which specifies the target set exactly. Let us assume for now that such a perfect rule exists. (Alternate versions of the rule might exist, but for our purposes, rules can be considered identical if they specify exactly the same set T.) The correct rule may be extremely complex, including conjunctions, disjunctions, and trade-offs among features. Your goal as a scientist, though, is to bring the hypothesized rule R_H in line with the correct rule R_T and thus to have the hypothesized set H match the target set T. You could then predict exactly which stars do and do not have planets. Similarly, a psychologist might wish to differentiate those who are at risk for schizophrenia from those who are not, or an epidemiologist might wish to understand who does and does not contract AIDS. The same structure can also be applied in a diagnostic context. For example, a diagnostician might seek to know the combination of signs that differentiates benign from malignant tumors.

In each case, an important component of the investigative process is the testing of hypotheses. That is, the investigator wants to know if the hypothesized rule R_H is identical to the correct rule R_T and if not, how they differ. This is accomplished through the collection of evidence, that is, the examination of instances. For example, you might choose a star hypothesized to have planets and train your telescope on it to see if it does indeed have planets, or you might examine tumors expected to be benign, to see if any are in fact malignant.

Wason (1960, 1968) developed a laboratory version of rule discovery to study people's hypothesis-testing strategies (in particular, their use of confirmation and disconfirmation), in a task that "simulates a miniature scientific problem" (1960, p. 139). In Wason's task, the universe was made up of all possible sets of three numbers ("triples"). Some of these triples fit the rule, in other words, conformed to a rule the experimenter had in mind. In our terms, fitting the experimenter's rule is the target property that subjects must learn to predict. The triples that fit the rule, then, constitute the target set, T. Subjects were provided with one target triple (2, 4, 6), and could ask the experimenter about any others they cared to. For each triple the subject proposed, the experimenter responded *yes* (fits the rule) or *no* (does not fit). Although subjects might start with only a vague guess, they quickly formed an initial hypothesis about the rule (R_H). For example, they might guess that the rule was "three consecutive even numbers." They could then perform one of two types of hypothesis tests (Htests): they could propose a triple they expected to be a target (e.g., 6, 8, 10), or a triple



Action	Result: "Yes" (in T)	Result: "No" (in $\bar{T}$)
+Htest	HnT: Ambiguous verification	Hn$\bar{T}$: Impossible
−Htest	$\bar{H}$nT: Conclusive falsification	$\bar{H}$n$\bar{T}$: Ambiguous verification

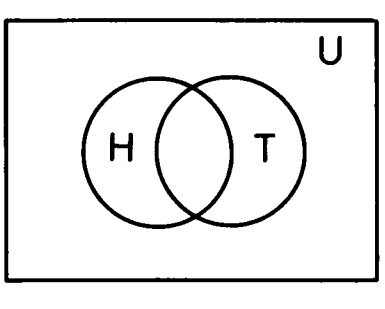
Figure 1. Representation of a situation in which the hypothesized rule is embedded within the correct rule, as in Wason's (1960) "2, 4, 6" task. (U = the universe of possible instances [e.g., all triples of numbers]; T = the set of instances that have the target property [e.g., they fit the experimenter's rule: increasing]; H = the set of instances that fit the hypothesized rule [e.g., increasing by 2].)

they expected not to be (e.g., 2, 4, 7). In this paper, we will refer to these as a positive hypothesis test (+Htest) and a negative hypothesis test (−Htest), respectively.

Wason found that people made much more use of +Htests than −Htests. The subject whose hypothesis was "consecutive evens," for example, would try many examples of consecutive-even triples and relatively few others. Subjects often became quite confident of their hypotheses after a series of +Htests only. In Wason's (1960) task this confidence was usually unfounded, for reasons we discuss later. Wason described the hypothesis testers as "seeking confirmation" because they looked predominantly at cases that fit their hypothesized rule for targets (e.g., different sets of consecutive even numbers). We think it more appropriate to view this "confirmation bias" as a manifestation of the general hypothesis-testing strategy we call the positive test (+ test) strategy. In rule discovery, the +test strategy leads to the predominant use of +Htests, in other words, a tendency to test cases you think will have the target property.

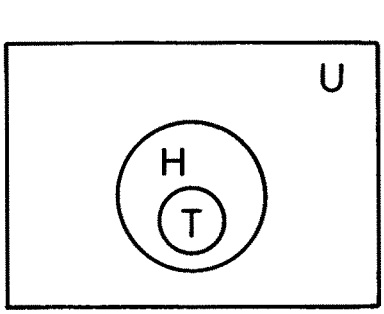
The general tendency toward +testing has been widely replicated. In a variety of different rule-discovery tasks (Klayman & Ha, 1985; Mahoney, 1976, 1979; Mynatt et al., 1977, 1978; Taplin, 1975; Tweney et al., 1980; Wason & Johnson-Laird, 1972) people look predominantly at cases they expect will have the target property, rather than cases they expect will not. As with nearly all strategies, people do not seem to adhere strictly to +testing, however. For instance, given an adequate number of test opportunities and a lack of pressure for a quick evaluation, people seem willing to test more widely (Gorman & Gorman, 1984; Klayman & Ha, 1985). Of particular interest is one manipulation that greatly improved success at Wason's 2, 4, 6 task. Tweney et al. (1980) used a task structurally identical to Wason's but modified the presentation of feedback. Triples were classified as either DAX or MED, rather than *yes* (fits the rule) or *no* (does not fit). The rule for DAX was Wason's original ascending-order rule, and all other triples were MED. Subjects in the DAX/MED version used even fewer −Htests than usual. However, they treated the DAX rule and the MED rule as two separate hypotheses, and tested each with +Htests, thereby facilitating a solution.

The thrust of this work has been more than just descriptive, however. There has been a strong emphasis on the notion that a +test strategy (or something like it) will lead to serious errors or inefficiencies in the testing of hypotheses. We begin by taking a closer look at this assumption. We examine what philosophers of science such as Popper and Platt have been arguing, and how that translates to prescriptions for information gathering in different hypothesis-testing situations. We then examine the task characteristics that control the extent to which a +test strategy deviates from those prescriptions. We begin with rule discovery as described above, and then consider what happens if additional information is available (examples of known targets and nontargets), and if an element of probabilistic error is introduced. The basic question is, if you are trying to determine



Action	Result: "Yes" (in T)	Result: "No" (in $\bar{T}$)
+Htest	HnT: Ambiguous verification	Hn$\bar{T}$: Conclusive falsification
−Htest	$\bar{H}$nT: Conclusive falsification	$\bar{H}$n$\bar{T}$: Ambiguous verification

Figure 2. Representation of a situation in which the hypothesized rule overlaps the correct rule.



Action	Result: "Yes" (in T)	Result: "No" (in $\bar{T}$)
+Htest	H∩T: Ambiguous verification	H∩$\bar{T}$: Conclusive falsification
−Htest	$\bar{H}$∩T: Impossible	$\bar{H}$∩$\bar{T}$: Ambiguous verification

Figure 3. Representation of a situation in which the hypothesized rule surrounds the correct rule.

the truth or falsity of a hypothesis, when is a +test strategy unwise and when is it not?

The Logic of Ambiguous Versus Conclusive Events

As a class, laboratory rule-discovery tasks share three simplifying assumptions. First, feedback is deterministically accurate. The experimenter provides the hypothesis tester with error-free feedback in accordance with an underlying rule. Second, the goal is to determine the one correct rule (R_T). All other rules are classified as incorrect, without regard to *how* wrong R_H may be, although the tester may be concerned with *where* it is wrong in order to form a new hypothesis. Third, correctness requires both sufficiency and necessity: A rule is incorrect if it predicts an instance will be in the target set when it is not (false positive), or predicts it will not be in the target set when it is (false negative). We discuss later the extent to which each of these assumptions restricts generalization to other tasks.

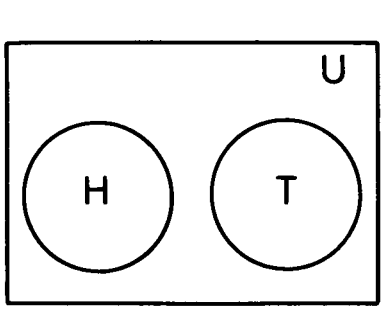
Consider again Wason's original task. Given the triple (2, 4, 6), the hypotheses that occur to most people are "consecutive even numbers," "increasing by 2," and the like. The correct rule, however, is much broader: "increasing numbers." Consider subjects whose hypothesized rule is "increasing by 2." Those who use only +Htests (triples that increase by 2, such as 6, 8, 10) can never discover that their rule is incorrect, because all examples of "increasing by 2" also fit the rule of "increasing." Thus, it is crucial to try −Htests (triples that do not increase by 2, such as 2, 4, 7). This situation is depicted in Figure 1. Here, U represents the universe of instances, all possible triples of numbers. T represents the target set, triples that fit the experimenter's rule ("increasing"). H represents the hypothesized set, triples that fit the tester's hypothesized rule (say, "increasing by 2"). There are in principle four classes of instances, although they do not all exist in this particular example:

1. H ∩ T: instances correctly hypothesized to be in the target set (positive hits).
2. H ∩ $\bar{T}$: instances incorrectly hypothesized to be in the target set (false positives).
3. $\bar{H}$ ∩ $\bar{T}$: instances correctly hypothesized to be outside the target set (negative hits).
4. $\bar{H}$ ∩ T: instances incorrectly hypothesized to be outside the target set (false negatives).

Instances of the types H ∩ $\bar{T}$ and $\bar{H}$ ∩ T *falsify* the hypothesis. That is, the occurrence of either shows conclusively that $H \neq T$, thus $R_H \neq R_T$; the hypothesized rule is not the correct one. Instances of the types H ∩ T and $\bar{H}$ ∩ $\bar{T}$ *verify* the hypothesis, in the sense of providing favorable evidence. However, these instances are ambiguous: The hypothesis may be correct, but these instances can occur even if the hypothesis is not correct. Note that there are only conclusive falsifications, no conclusive verifications. This logical condition is the backbone of philosophies of science that urge investigators to seek falsification rather than verification of their hypotheses (e.g., Popper, 1959). Put somewhat simplistically, a lifetime of verifications can be countered by a single conclusive falsification, so it makes sense for scientists to make the discovery of falsifications their primary goal.

Suppose, then, that you are the tester in Wason's task, with the hypothesis of "increasing by 2." If you try a +Htest (e.g., 6, 8, 10) you will get either a *yes* response, which is an ambiguous verification of the type H ∩ T, or a *no*, which is a conclusive falsification of the type H ∩ $\bar{T}$. The falsification H ∩ $\bar{T}$ would show that meeting the conditions of your rule is not sufficient to guarantee membership in T. Thus, +Htests can be said to be tests of the rule's sufficiency. However, unknown to the subjects in the 2, 4, 6, task (Figure 1) there are no instances of H ∩ $\bar{T}$, because the hypothesized rule is sufficient: Any instance following R_H ("increasing by 2") will in fact be in the target set T ("increasing"). Thus, +Htests will never produce falsification. If you instead try a −Htest (e.g., 2, 4, 7) you will get either a *no* answer which is an ambiguous verification ($\bar{H}$ ∩ $\bar{T}$) or a *yes* answer which is a conclusive falsification ($\bar{H}$ ∩ T). The falsification $\bar{H}$ ∩ T shows that your conditions are not necessary for membership in T. Thus, −Htests test a rule's necessity. In the 2, 4, 6 task, −Htests can result in conclusive falsification because R_H is sufficient but not necessary (i.e., there are some target triples that do not increase by 2).

In the above situation, the Popperian exhortation to seek falsification can be fulfilled only by −Htesting, and those who rely on +Htests are likely to be misled by the abundant verification they receive. Indeed, Wason deliberately designed his task so that this would be the case, in order to show the pitfalls of "confirmation bias" (Wason, 1962). The hypothesis-tester's situation is not always like this, however. Consider the situation in which the hypothesized set merely overlaps the target set, as shown in Figure 2, rather than being *embedded* within it, as



Action	Result: "Yes" (in T)	Result: "No" (in $\bar{T}$)
+Htest	H∩T: Impossible	H∩$\bar{T}$: Conclusive falsification
−Htest	$\bar{H}$∩T: Conclusive falsification	$\bar{H}$∩$\bar{T}$: Ambiguous verification

Figure 4. Representation of a situation in which the hypothesized rule and the correct rule are disjoint.

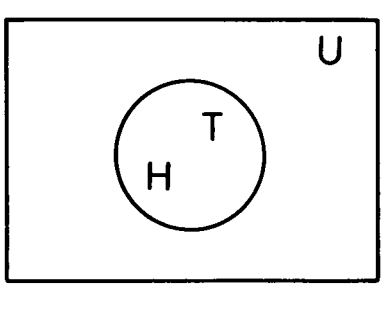
shown in Figure 1. This would be the case if, for example, the correct rule were "three even numbers." There would be some members of H ∩ $\bar{T}$, instances that were "increasing by 2" but not "three evens" (e.g., 1, 3, 5), and some members of $\bar{H}$ ∩ T, "three evens" but not "increasing by 2" (e.g., 4, 6, 2). Thus, conclusive falsification could occur with either +Htests or −Htests. Indeed, it is possible to be in a situation just the opposite of Wason's, shown in Figure 3. Here, the hypothesis is too broad and "surrounds" the target set. This would be the case if the correct rule were, say, "consecutive even numbers." Now a tester who did only *−Htests* could be sorely misled, because there are no falsifications of the type $\bar{H}$ ∩ T; any instance that violates "increasing by 2" also violates "consecutive evens." Only *+Htests* can reveal conclusive falsifications (H ∩ $\bar{T}$ instances such as 1, 3, 5).

Aside from these three situations, there are two other possible relationships between H and T. When H and T are disjoint (Figure 4), any +Htest will produce conclusive falsification, because nothing in H is in T; −Htests could produce either verification or falsification. This is not likely in the 2, 4, 6 task, because you are given one known target instance to begin with. In the last case (Figure 5), you have finally found the correct rule, and H coincides with T. Here, every test produces ambiguous information; a final proof is possible only if there is a finite universe of instances and every case is searched.

In naturally occurring situations, as in Wason's (1960) task, one could find oneself in any of the conditions depicted, usually with no way of knowing which. Suppose, for example, that you are a manufacturer trying to determine the best way to advertise your line of products, and your current hypothesis is that television commercials are the method of choice. For you, the universe, U, is the set of possible advertising methods; the target set, T, is the set of methods that are effective, and the hypothesized set, H, is television commercials. Suppose that in fact the set of effective advertising methods for these products is much broader: any visual medium (magazine ads, etc.) will work. This is the situation depicted in Figure 1. If you try +Htests (i.e., try instances in your hypothesized set, television commercials) you will never discover that your rule is wrong, because television commercials will be effective. Only by trying things you think will not work (−Htests) can you obtain falsification. You might then discover an instance of the type $\bar{H}$ ∩ T: nontelevision advertising that is effective.

Suppose instead that the correct rule for effectively advertising these products is to use humor. This is the situation in Figure 2. You could find a (serious) television commercial that you thought would work, but does not (H ∩ $\bar{T}$), or a (humorous) nontelevision ad that you thought would not work, but does ($\bar{H}$ ∩ T). Thus, conclusive falsification could occur with either a +Htest or a −Htest. If instead the correct rule for these products is more restricted, say, "prime-time television only," you would have an overly broad hypothesis, as shown in Figure 3. In that case, you will never obtain falsification if you use −Htests (i.e., if you experiment with methods you think will not work), because anything that is not on television is also not on prime time. Only +Htests can reveal conclusive falsifications, by finding instances of H ∩ $\bar{T}$ (instances of television commercials that are not effective).

What is critical, then, is not the testing of cases that do not fit your hypothesis, but the testing of cases that are most likely



Action	Result: "Yes" (in T)	Result: "No" (in $\bar{T}$)
+Htest	H∩T: Ambiguous verification	H∩$\bar{T}$: Impossible
−Htest	$\bar{H}$∩T: Impossible	$\bar{H}$∩$\bar{T}$: Ambiguous verification

Figure 5. Representation of the situation in which the hypothesized rule coincides with the correct rule.

to prove you wrong. In Wason's task these two actions are identical, but as shown in Figures 2 through 5, this is not generally so. Thus, it is very important to distinguish between two different senses of "seeking disconfirmation." One sense is to examine instances that you predict will not have the target property. The other sense is to examine instances you most expect to falsify, rather than verify, your hypothesis. This distinction has not been well recognized in past analyses, and confusion between the two senses of disconfirmation has figured in at least two published debates, one involving Wason (1960, 1962) and Wetherick (1962), the other involving Mahoney (1979, 1980), Hardin (1980), and Tweney, Doherty, and Mynatt (1982). The prescriptions of Popper and Platt emphasize the importance of falsification of the hypothesis, whereas empirical investigations have focused more on the testing of instances outside the hypothesized set.

Confirmation and Disconfirmation: Where's the Information?

The distinction between −testing and seeking falsification leads to an important question for hypothesis testers: Given the choice between +tests and −tests, which is more likely to yield critical falsification? As is illustrated in Figures 1 through 5, the answer depends on the relation between your hypothesized set and the target set. This, of course, is impossible to know without first knowing what the target set is. Even without prescience of the truth, however, it is possible for a tester to make a reasoned judgment about which kind of test to perform. Prescriptions can be based on (at least) two considerations: (a) What type of errors are of most concern, and (b) Which test could be expected, probabilistically, to yield conclusive falsification more often. The first point hinges on the fact that +Htests and −Htests reveal different kinds of errors (false positives and false negatives, respectively). A tester might care more about one than the other and might be advised to test accordingly. Although there is almost always some cost to either type of error, one cost may be much higher than the other. For example, a personnel director may be much more concerned about hiring an incompetent person ($H \cap \bar{T}$) than about passing over some potentially competent ones ($\bar{H} \cap T$). Someone in this position should favor +Htests (examining applicants judged competent, to find any failures) because they reveal potential false positives. On the other hand, some situations require greater concern with false negatives than false positives. For example, when dealing with a major communicable disease, it is more serious to allow a true case to go undiagnosed and untreated ($\bar{H} \cap T$) than it is to mistakenly treat someone ($H \cap \bar{T}$). Here the emphasis should be on −Htests (examining people who test negative, to find any missed cases), because they reveal potential false negatives.

It could be, then, that a preference for +Htests merely reflects a greater concern with sufficiency than necessity. That is, the tester may simply be more concerned that all chosen cases are true than that all true cases are chosen. For example, experiments by Vogel and Annau (1973), Tschirgi (1980), and Schwartz (1981, 1982) suggest that an emphasis on the sufficiency of one's actions is enhanced when one is rewarded for each individual success rather than only for the final rule discovery. Certainly, in many real situations (choosing an employee, a job, a spouse, or a car) people must similarly live with their mistakes. Thus, people may be naturally inclined to focus more on false positives than on false negatives in many situations. A tendency toward +Htesting would be entirely consistent with such an emphasis. However, it is still possible that people retain an emphasis on sufficiency when it is inappropriate (as in Wason's task).

Suppose that you are a tester who cares about both sufficiency and necessity: your goal is simply to determine whether or not you have found the correct rule. It is still possible to analyze the situation on the basis of reasonable expectations about the world. If you accept the reasoning of Popper and Platt, the goal of your testing should be to uncover conclusive falsifications. Which kind of test, then, should you expect to be more likely to do so? Assume that you do not know in advance whether your hypothesized set is embedded in, overlaps, or surrounds the target. The general case can be characterized by four quantities[1]:

- $p(t)$ The overall base-rate probability that a member of the domain is in the target set. This would be, for example, the proportion of stars in the galaxy that have planets.
- $p(h)$ The overall probability that a member of the domain is in the hypothesized set. This would be the proportion of stars that fit your hypothesized criteria for having planets.
- $z^+ = p(\bar{t}|h)$ The overall probability that a positive prediction will prove false, for example, that a star hypothesized to have planets will turn out not to.
- $z^- = p(t|\bar{h})$ The overall probability that a negative prediction will prove false, for example, that a star hypothesized not to have planets will turn out in fact to have them.

The quantities z^+ and z^- are indexes of the errors made by the hypothesis. They correspond to the false-positive rate and false-negative rate for the hypothesized rule R_H (cf. Einhorn & Hogarth, 1978). In our analyses, all four of the above probabilities are assumed to be greater than zero but less than one.[2] This corresponds to the case of overlapping target and hypothesis sets, as shown in Figure 2. However, other situations can be regarded as boundary conditions to this general case. For example, the embedded, surrounding, and coincident situations (Figures 1, 3, and 5) are cases in which $z^+ = p(\bar{t}|h) = 0$, $z^- = p(t|\bar{h}) = 0$, or both, respectively, and in the disjoint situation (Figure 4), $z^+ = 1$.

Recall that there are two sets of conclusive falsifications: $H \cap \bar{T}$ (your hypothesis predicts planets, but there are none), and $\bar{H} \cap T$ (your hypothesis predicts no planets, but there are some). If you perform a +Htest, the probability of a conclusive falsification, $p(Fn|+Htest)$, is equal to the false positive rate, z^+. If you perform a −Htest, the chance of falsification,

[1] We use a lowercase letter to designate an instance of a given type: t is an instance in set T, $\bar{t}$ is an instance in $\bar{T}$, and so on.

[2] Our analyses treat the sets U, T, and H as finite, but also apply to infinite sets, as long as T and H designate finite, nonzero fractions of U. In Wason's task (1960), for example, if U = all sets of three numbers and H = all sets of three even numbers, then we can say that H designates 1/8 of all the members of U, in other words, $p(h) = 1/8$.

$p(\text{Fn}|-\text{Htest})$, is equal to the false negative rate, z^-. A Popperian hypothesis-tester might wish to perform the type of test with the higher expected chance of falsification. Of course, you cannot have any direct evidence on z^+ and z^- without obtaining some falsification, at which point you would presumably form a different hypothesis. However, the choice between tests does not depend on the values of z^+ and z^- per se, but on the relationship between them, and that is a function of two quantities about which an investigator might well have some information: $p(\text{t})$ and $p(\text{h})$. What is required is an estimate of the base rate of the phenomenon you are trying to predict (e.g., what proportion of stars have planets, what proportion of the population falls victim to schizophrenia or AIDS, what proportion of tumors are malignant) and an estimate of the proportion your hypothesis would predict. Then

$$\begin{aligned} z^+ = p(\bar{\text{t}}|\text{h}) &= 1 - p(\text{t}|\text{h}) \\ &= 1 - p(\text{t} \cap \text{h})/p(\text{h}) \\ &= 1 - [p(\text{t}) - p(\text{t} \cap \bar{\text{h}})]/p(\text{h}) \\ &= 1 - \frac{p(\text{t})}{p(\text{h})} + \frac{p(\text{t}|\bar{\text{h}}) \cdot p(\bar{\text{h}})}{p(\text{h})} \\ z^+ &= z^- \cdot \frac{p(\bar{\text{h}})}{p(\text{h})} + \left(1 - \frac{p(\text{t})}{p(\text{h})}\right). \end{aligned} \tag{1}$$

According to Equation 1, even if you have no information about z^+ and z^-, you can estimate their relationship from estimates of the target and hypothesis base rates, $p(\text{t})$ and $p(\text{h})$. It is not necessarily the case that the tester knows these quantities exactly. However, there is usually some evidence available for forming estimates on which to base a judgment. In any case, it is usually easier to estimate, say, how many people suffer from schizophrenia than it is to determine the conditions that produce it.

It seems reasonable to assume that in many cases the tester's hypothesis is at least about the right size. People are not likely to put much stock in a hypothesis that they believe greatly overpredicts or underpredicts the target phenomenon. Let us assume, then, that you believe that $p(\text{h}) \approx p(\text{t})$. Under these circumstances, Equation 1 can be approximated as

$$z^+ = \frac{p(\bar{\text{t}})}{p(\text{t})} \cdot z^- \tag{2}$$

Thus, if $p(\text{t}) < .5$, then $z^+ > z^-$, which means that $p(\text{Fn}|+\text{Htest}) > p(\text{Fn}|-\text{Htest})$. In other words, if you are attempting to predict a minority phenomenon, you are more likely to receive falsification using +Htests than −Htests. We would argue that, in fact, real-world hypothesis testing most often concerns minority phenomena. For example, a recent estimate for the proportion of stars with planets is ⅓ (Sagan, 1980, p. 300), for the prevalence of schizophrenia, less than 1% (American Psychiatric Association, 1980), and for the incidence of AIDS in the United States, something between 10^{-4} and 10^{-5} (Centers for Disease Control, 1986). Even in Wason's original task (1960), the rule that seemed so broad (any increasing) has a $p(\text{t})$ of only 1/6, assuming one chooses from a large range of numbers. Indeed, if $p(\text{t})$ were greater than .5, the perception of target and nontarget would likely reverse. If 80% of the population had some disease, immunity would be the target property, and $p(\text{t})$ would then be .2 (cf. Bourne & Guy, 1968; Einhorn & Hogarth, 1986).

Table 1
Conditions Favoring +Htests or −Htests as Means of Obtaining Conclusive Falsification

Target and hypothesis base rates	Comparison of probability of falsification (Fn) for +Htests and −Htests[a]
$p(\text{t}) < .5$	
$p(\text{t}) > p(\text{h})$	Depends on specific values of z^+ and z^-
$p(\text{t}) = p(\text{h})$	$p(\text{Fn}|+\text{Htest}) > p(\text{Fn}|-\text{Htest})$
$p(\text{t}) < p(\text{h}) \leq .5$	$p(\text{Fn}|+\text{Htest}) > p(\text{Fn}|-\text{Htest})$
$p(\text{t}) < .5 < p(\text{h})$	Depends on specific values of z^+ and z^-
$p(\text{t}) \geq .5$	
$p(\text{t}) \geq .5 > p(\text{h})$	Depends on specific values of z^+ and z^-
$p(\text{t}) > p(\text{h}) \geq .5$	$p(\text{Fn}|+\text{Htest}) < p(\text{Fn}|-\text{Htest})$
$p(\text{t}) = p(\text{h})$	$p(\text{Fn}|+\text{Htest}) \leq p(\text{Fn}|-\text{Htest})$
$p(\text{t}) < p(\text{h})$	Depends on specific values of z^+ and z^-

[a] See Equation 1 for derivation.

Thus, under some very common conditions, the probability of receiving falsification with +Htests could be much greater than with −Htests. Intuitively, this makes sense. When you are investigating a relatively rare phenomenon, $p(\text{t})$ is low and the set $\bar{\text{H}}$ is large. Finding a t in $\bar{\text{H}}$ (obtaining falsification with −Htests) can be likened to the proverbial search for a needle in a haystack. Imagine, for example, looking for AIDS victims among people believed not at risk for AIDS. On the other hand, these same conditions also mean that $p(\bar{\text{t}})$ is high, and set H is small. Thus, finding a $\bar{\text{t}}$ in H (with +Htests) is likely to be much easier. Here, you would be examining people with the hypothesized risk factors. If you have a fairly good hypothesis, $p(\bar{\text{t}}|\text{h})$ is appreciably lower than $p(\bar{\text{t}})$, but you are still likely to find healthy people in the hypothesized risk group, and these cases are informative. (You might also follow a strategy based on examining *known* victims; we discuss this kind of testing later.)

The conditions we assume above (a minority phenomenon, and a hypothesis of about the right size) seem to apply to many naturally occurring situations. However, these assumptions may not always hold. There may be cases in which a majority phenomenon is the target (e.g., because it was unexpected); then $p(\text{t}) > .5$. There may also be situations in which a hypothesis is tested even though it is not believed to be the right size, so that $p(\text{h}) \neq p(\text{t})$. For example, you may not be confident of your estimate for either $p(\text{t})$ or $p(\text{h})$, so you are not willing to reject a theoretically appealing hypothesis on the basis of those estimates. Or you may simply not know what to add to or subtract from your hypothesis, so that a search for falsification is necessary to suggest where to make the necessary change. In any case, a tester with some sense of the base rate of the phenomenon can make a reasoned guess as to which kind of test is more powerful, in the sense of being more likely to find critical falsification. The conditions under which +Htests or −Htests are favored are summarized in Table 1.

There are two main conclusions to be drawn from this analysis. First, it is important to distinguish between two possible senses of "seeking disconfirmation": (a) testing cases your hy-

Table 2
Conditions Favoring +Ttests or −Ttests as Means of Obtaining Conclusive Falsification

Target and hypothesis base rates	Comparison of probability of falsification (Fn) for +Ttests and −Ttests[a]
$p(t) < .5$	
$p(t) > p(h)$	$p(Fn\|+Ttest) > p(Fn\|-Ttest)$
$p(t) = p(h)$	$p(Fn\|+Ttest) > p(Fn\|-Ttest)$
$p(t) < p(h)$	Depends on specific values of x^+ and x^-
$p(t) \geq .5$	
$p(t) > p(h)$	Depends on specific values of x^+ and x^-
$p(t) = p(h)$	$p(Fn\|+Ttest) \leq p(Fn\|-Ttest)$
$p(t) < p(h)$	$p(Fn\|+Ttest) < p(Fn\|-Ttest)$

[a] See Equation 3 for derivation.

pothesis predicts to be nontargets, and (b) testing cases that are most likely to falsify the hypothesis. It is the latter that is generally prescribed as optimal. Second, the relation between these two actions depends on the structure of the environment. Under some seemingly common conditions, the two actions can, in fact, conflict. The upshot is that, despite its shortcomings, the +test strategy may be a reasonable way to test a hypothesis in many situations. This is not to say that human hypothesis testers are actually aware of the task conditions that favor or disfavor the use of a +test strategy. Indeed, people may not be aware of these factors precisely because the general heuristic they use often works well.

Information in Target Tests

The 2, 4, 6 task involves only one-half of the proposed +test strategy, that is, the testing of cases hypothesized to have the target property (+Htesting). In some tasks, however, the tester may also have an opportunity to examine cases in which the target property is *known* to be present (or absent) and to receive feedback about whether the instance fits the hypothesis. For example, suppose that you hypothesize that a certain combination of home environment, genetic conditions, and physical health distinguishes schizophrenic individuals from others. It would be natural to select someone diagnosed as schizophrenic and check whether the hypothesized conditions were present. We will call this a positive target test (+Ttest), because you select an instance known to be in the target set. Similarly, you could examine the history of someone judged not to be schizophrenic to see if the hypothesized conditions were present. We call this a negative target test (−Ttest). Generally, Ttests may be more natural in cases involving diagnostic or epidemiological questions, when one is faced with known effects for which the causes and correlates must be determined.

Ttests behave in a manner quite parallel to the Htests described above. A +Ttest results in verification ($T \cap H$) if the known target turns out to fit the hypothesized rule (e.g., someone diagnosed as schizophrenic turns out to have the history hypothesized to be distinctive to schizophrenia). A +Ttest results in falsification if a known target fails to have the features hypothesized to distinguish targets ($T \cap \bar{H}$). The probability of falsification with a +Test, designated x^+, is $p(\bar{h}|t)$. This is equivalent to the miss rate of signal detection theory (Green & Swets, 1966). The falsifying instances revealed by +Ttests (missed targets, $T \cap \bar{H}$) are the same kind revealed by −Htests (false negatives, $\bar{H} \cap T$). Note, though, that the miss rate of +Ttests is calculated differently than the false negative rate of −Htests [$x^+ = p(\bar{h}|t)$; $z^- = p(t|\bar{h})$]. Both +Ttests and −Htests assess whether the conditions in R_H are *necessary* for schizophrenia.

With −Ttests, verifications are of the type $\bar{T} \cap \bar{H}$ (nonschizophrenics who do not have the history hypothesized for schizophrenics), and falsifications are of the type $\bar{T} \cap H$ (nonschizophrenics who do have that history). The probability of falsification with −Ttests, designated x^-, is $p(h|\bar{t})$. This is equivalent to the false alarm rate in signal detection theory. −Ttests and +Htests reveal the same kinds of falsifying instances (false alarms or false positives). The rate of falsification with −Ttests is $x^- = p(h|\bar{t})$ compared to $z^+ = p(\bar{t}|h)$ for +Htests. Both −Ttests and +Htests assess whether the conditions in R_H are *sufficient*.

We can compare the two types of Ttests in a manner parallel to that used to compare Htests. The values x^+ and x^- (the miss rate and false alarm rate, respectively) can be related following the same logic used in Equation 1:

$$x^+ = x^- \frac{p(\bar{t})}{p(t)} + \left(1 - \frac{p(h)}{p(t)}\right). \tag{3}$$

If we again assume that $p(t) < .5$ and $p(h) = p(t)$, then $x^+ > x^-$. This means that +Ttests are more likely to result in falsification than are −Ttests. The full set of conditions favoring one type of Ttest over the other are shown in Table 2. Under common circumstances, it can be normatively appropriate to have a second kind of "confirmation bias," namely, a tendency to test cases known to be targets rather than those known to be nontargets.

It is also interesting to consider the relations between Ttests and Htests. In some situations, it may be more natural to think about one or the other. In an epidemiological study, for example, cases often come presorted as T or $\bar{T}$ (e.g., diagnosed victims of disease vs. normal individuals). In an experimental study, on the other hand, the investigator usually determines the presence or absence of hypothesized factors and thus membership in H or $\bar{H}$ (e.g., treatment vs. control group). Suppose, though, that you are in a situation where all four types of test are feasible. There are then two tests that reveal falsifications of the type $H \cap \bar{T}$ (false positives or false alarms), namely +Htests and −Ttests. These falsifications indicate that the hypothesized conditions are not *sufficient* for the target phenomenon. For example, suppose a team of meteorologists wants to test whether certain weather conditions are sufficient to produce tornadoes. The team can look for tornadoes where the hypothesized conditions exist (+Htests) or they can test for the conditions where tornadoes have not occurred (−Ttests). The probability of discovering falsification with each kind of test is as follows:

$$p(Fn|+Htest) = z^+ = p(\bar{t}|h) = \frac{p(h \cap \bar{t})}{p(h)}$$

$$p(Fn|-Ttest) = x^- = p(h|\bar{t}) = \frac{p(h \cap \bar{t})}{p(\bar{t})}$$

$$z^+ = x^- \cdot \frac{p(\bar{t})}{p(h)}. \quad (4)$$

Thus, if we assume, as before, that $p(t) < .5$, and $p(h) = p(t)$, then $z^+ > x^-$: the probability of finding a falsifying instance ($h \cap \bar{t}$) is higher with +Htests than with −Ttests.

There are also two tests that reveal falsifications of the type $\bar{H} \cap T$ (false negatives or misses): +Ttests and −Htests. These falsifications indicate that the hypothesized conditions are not *necessary* for the target phenomenon. The meteorologists can test whether the hypothesized weather conditions are necessary for tornadoes by looking at conditions where tornadoes are sighted (+Ttests) or by looking for tornadoes where the hypothesized conditions are lacking (−Htests). The probability of falsification with these two tests can be compared, parallel to Equation 4, above:

$$x^+ = z^- \cdot \frac{p(\bar{h})}{p(t)}. \quad (5)$$

Thus, the probability of finding $\bar{H} \cap T$ falsifications is higher with +Ttests than with −Htests.

These relationships reinforce the idea that it may well be advantageous in many situations to have two kinds of "confirmation bias" in choosing tests: a tendency to examine cases hypothesized to be targets (+Htests) and a tendency to examine cases known to be targets (+Ttests). Taken together, these two tendencies compose the general +test strategy. Under the usual assumptions [$p(t) < .5$ and $p(t) \approx p(h)$], +Htests are favored over −Htests, and +Ttests over −Ttests, as more likely to find falsifications. Moreover, if you wish to test your rule's sufficiency, +Htests are better than −Ttests; if you wish to test the rule's necessity, +Ttests are better than −Htests. Thus, it may be advantageous for the meteorologists to focus their field research on areas with hypothesized tornado conditions and areas of actual tornado sighting (which, in fact, they seem to do; see Lucas & Whittemore, 1985). Like many other cognitive heuristics, however, this +test heuristic may prove maladaptive in particular situations, and people may continue to use the strategy in those situations nonetheless (cf. Hogarth, 1981; Tversky & Kahneman, 1974).

Hypothesis Testing in Probabilistic Environments

Laboratory versions of rule discovery usually take place in a deterministic environment: There is a correct rule that makes absolutely no errors, and feedback about predictions is completely error-free (see Kern, 1983, and Gorman, 1986, for interesting exceptions). In real inquiry, however, one does not expect to find a rule that predicts every schizophrenic individual or planetary system without error, and one recognizes that the ability to detect psychological disorders or celestial phenomena is imperfect. What, then, is the normative status of the +test heuristic in a probabilistic setting?

Irreducible error. In a probabilistic environment, it is somewhat of a misnomer to call any hypothesis correct, because even the best possible hypothesis will make some false-positive and false-negative predictions. These irreducible errors might actually be due to imperfect feedback, but from the tester's point of view they look like false positives or false negatives. Alternatively, the world may have a truly random component, or the problem may be so complex that in practice perfect prediction would be beyond human reach. In any case, the set T can be defined as the set of instances that the feedback indicates are targets. A best possible rule, R_B, can be postulated that defines the set B. B matches T as closely as possible, but not exactly. Because of probabilistic error, even the best rule makes false-positive and false-negative prediction errors (i.e., $p(\bar{t}|b) > 0$ and $p(t|\bar{b}) > 0$). The probabilities of these errors, designated ϵ^+ and ϵ^-, represent theoretical or practical minimum error rates.[3]

Qualitatively, the most important difference between deterministic and probabilistic environments is that both verification and falsification are of finite value and subject to some degree of probabilistic error. Thus, falsifications are not conclusive but merely constitute some evidence against the hypothesis, and verifications must also be considered informative, despite their logical ambiguity. Ultimately, it can never be known with certainty that any given hypothesis is or is not the best possible. One can only form a belief about the probability that a given hypothesis is correct, in light of the collected evidence.

Despite these new considerations, it can be shown that the basic findings of our earlier analyses still apply. Although the relationship is more complicated, the relative value of +tests and −tests is still a function of estimable task characteristics. In general, it is still the case that +tests are favored when $p(t)$ is small and $p(h) \approx p(t)$, as suggested earlier. Although we discuss only Htests here, a parallel analysis can be performed for Ttests as well.

Revision of beliefs. Assume that your goal is to obtain the most evidence you can about whether or not your current hypothesis is the best possible. Which type of test will, on average, be more informative? This kind of problem calls for an analysis of the expected value of information (e.g., see Edwards, 1965; Raiffa, 1968). Such analyses are based on Bayes's equation, which provides a normative statistical method for assessing the extent to which a subjective degree of belief should be revised in light of new data. To perform a full-fledged Bayesian analysis of value of information, it would be necessary to represent the complete reward structure of the particular task and compute the tester's subjective expected utility of each possible action. Such an analysis would be very complex or would require a great many simplifying assumptions. It is possible, though, to use a simple, general measure of "impact," such as the expected change in belief (EΔP).

Suppose you think that there is some chance your hypothesis is the best possible, $p(R_H = R_B)$. Then, you perform a +Htest, and receive a verification (Vn). You would now have a somewhat higher estimate of the chance that your hypothesis is the best one $p(R_H = R_B|Vn, +H)$. Call the impact of this test $\Delta P_{Vn,+H}$, the absolute magnitude of change in degree of belief. Of course, you might have received a falsification (Fn) instead, in which case your belief that $R_H = R_B$ would be reduced by some amount, $\Delta P_{Fn,+H}$. The expected change in belief for a +Htest,

[3] For simplicity, we ignore the possibility that a rule might produce, say, fewer false positives but more false negatives than the best rule. We assume that the minimum ϵ^+ and ϵ^- can both be achieved at the same time. The more general case could be analyzed by defining a joint function of ϵ^+ and ϵ^- which is to be minimized.

given that you do not know in advance whether you will receive a verification or a falsification, would thus be

$$E\Delta P_{+H} = p(\text{Fn}|+\text{Htest}) \cdot \Delta P_{\text{Fn},+H} + p(\text{Vn}|+\text{Htest}) \cdot \Delta P_{\text{Vn},+H}. \quad (6)$$

In the appendix, we show that

$$\Delta P_{\text{Fn},+H} = 1 - \frac{\epsilon^+}{z^+}, \quad (7)$$

$$\Delta P_{\text{Vn},+H} = \frac{1 - \epsilon^+}{1 - z^+} - 1, \quad (8)$$

$$p(\text{Fn}|+\text{Htest}) \cdot \Delta P_{\text{Fn},+H} = (z^+ - \epsilon^+) \cdot p(R_H = R_B), \quad (9)$$

and

$$p(\text{Vn}|+\text{Htest}) \cdot \Delta P_{\text{Vn},+H} = (z^+ - \epsilon^+) \cdot p(R_H = R_B). \quad (10)$$

Thus

$$E\Delta P_{+H} = 2(z^+ - \epsilon^+) \cdot p(R_H = R_B). \quad (11)$$

Similarly,

$$E\Delta P_{-H} = 2(z^- - \epsilon^-) \cdot p(R_H = R_B). \quad (12)$$

This probabilistic analysis looks different from its deterministic counterpart in one respect. Before, the emphasis was strictly on falsification. Here, verification can sometimes be more informative than falsification. Using +Htests to illustrate, Equations 7 and 8 imply that if $z^+ > .5$, then $\Delta P_{\text{Vn},+H} > \Delta P_{\text{Fn},+H}$. A hypothesis with $z^+ > .5$ is a weak hypothesis; you believe the majority of predicted targets will prove wrong. Perhaps this is an old hypothesis that is now out of favor, or a new shot-in-the-dark guess. The ΔP measure captures the intuition that surprise verification of a longshot hypothesis has more impact than the anticipated falsification.

In considering the expected impact of a test, you must balance the greater impact of unexpected results against the fact that you do not think such results are likely to happen. With the $E\Delta P$ measure, the net result is that verifications and falsifications are expected to make equal contributions to changes in belief, overall (as shown in Equations 9 and 10). Verifications and falsifications have equal expected impact even in a deterministic environment, according to this definition of impact. The deterministic environment is merely a special case in which $\epsilon^+ = \epsilon^- = 0$.

Given this probabilistic view of the value of verification and falsification, where should one look for information? The answer to this question, based on the comparison between +Htests and −Htests, changes very little from the deterministic case. It would be a rational policy for a tester to choose the type of Htest associated with the greatest expected change in belief. In that case, according to Equations 11 and 12, you want to choose the test for which $z - \epsilon$ is greatest: +Htests if $(z^+ - \epsilon^+) > (z^- - \epsilon^-)$. In other words, choose the test for which you believe the probability of falsification (z) is most above the level of irreducible error (ϵ). This prescription is obviously very similar to the conditions specified for the deterministic environment. Indeed, if the two ϵs are equal (even if nonzero) the rule is identical: Choose the test with the higher z. Thus, the prescriptions shown in Table 1 hold in a probabilistic environment, as long as irreducible error is also taken into account. In the Appendix we also present an alternative measure of informativeness (a measure of "diagnosticity" often used in Bayesian analyses); the basic premises of our comparison remain intact. Qualitatively similar results obtain even when using a non-Bayesian analysis, based on statistical information theory (see Klayman, 1986).

Information in Hypothesis Testing: Conclusions

The foundation of our analysis is the separation of disconfirmation as a goal from disconfirmation as a search strategy. It is a widely accepted prescription that an investigator should seek falsification of hypotheses. Our analyses show, though, that there is no correspondingly simple prescription for the search strategy best suited to that goal. The optimal strategy is a function of a variety of task variables such as the base rates of the target phenomenon and the hypothesized conditions. Indeed, even attempting falsification is not necessarily the path to maximum information (see also Klayman, 1986).

We do not assume that people are aware of the task variables that determine the best test strategies. Rather, we suggest that people use a general, all-purpose heuristic, the positive test strategy, which is applied across a broad range of hypothesis-testing tasks. Like any all-purpose heuristic, this +test strategy is not always optimal and can lead to serious difficulties in certain situations (as in Wason's 2, 4, 6 task). However, our analyses show that +testing is not a bad approach in general. Under commonly occurring conditions, the +test strategy leads people to perform tests of both sufficiency and necessity (+Htests and +Ttests), using the types of tests most likely to discover violations of either.

Beyond Rule Discovery: The Positive Test Strategy in Other Contexts

The main point of our analysis is not that people are better hypothesis testers than previously thought (although that may be so). Rather, the +test strategy can provide a basis for understanding the successes and failures of human hypothesis testing in a variety of situations. In this section, we apply our approach to several different hypothesis-testing situations. Each of the tasks we discuss has an extensive research literature of its own. However, there has been little cross-task generality beyond the use of the common "confirmation bias" label. We show how these diverse tasks can be given an integrative interpretation based on the general +test strategy. Each task has its unique requirements, and ideally, people should adapt their strategies to the characteristics of the specific task at hand. People may indeed respond appropriately to some of these characteristics under favorable conditions (when there is concrete task-specific information, light memory load, adequate time, extensive experience, etc.). We propose that, under less friendly conditions, hypothesis testers rely on a generally applicable default approach based on the +test strategy.

Concept Identification

At the beginning of this paper, we described the concept-identification task (Bruner et al., 1956) as a forerunner of Wason's rule-discovery task (Wason, 1960). In both tasks, the subject's goal is to identify the rule or concept that determines which of a subset of stimuli are designated as correct. In concept identification, however, the set of possible instances and possible rules

is highly restricted. For example, the stimuli may consist of all combinations of four binary cues (letter X or T, large or small, black or white, on the right or left), with instructions to consider only simple (one-feature) rules (e.g., Levine, 1966). The hypothesis set, then, is restricted to only eight possibilities. Even when conjunctions or disjunctions of features are allowed (e.g., Bourne, 1974; Bruner et al., 1956), the hypothesis set remains circumscribed.

A number of studies of concept identification have documented a basic win-stay, lose-shift strategy (e.g., see Levine, 1966, 1970; Trabasso & Bower, 1968). That is, the learner forms an initial hypothesis about which stimuli are reinforced (e.g., "Xs on the left") and responds in accordance with that hypothesis as long as correct choices are produced. If an incorrect choice occurs, the learner shifts to a new hypothesis and responds in accordance with that, and so on. In our terms, this is +Htesting. It is what we would expect to see, especially since total success requires a rule that is sufficient for reward, only. In the concept-identification task +Htesting alone could lead to a successful solution. However, because there are only a finite number of instances (cue combinations), and a finite number of hypotheses, +testing is not the most effective strategy. A more efficient strategy is to partition the hypotheses into classes and perform a test that will eliminate an entire class of hypotheses in a single trial. For example, if a small, black X on the left is correct on one trial, the rules "large," "white," "T," and "right" can all be eliminated at once. If on the next trial a large, black X on the right is correct, only "black" and "X" remain as possibilities, ignoring combinations. This "focusing" strategy (Bruner et al., 1956) is mathematically optimal but requires two things from subjects. First, they must recognize that having a circumscribed hypothesis set means it is possible to use a special efficient strategy not otherwise available. Second, focusing requires considerable cognitive effort to design an efficient sequence of tests and considerable memory demands to keep track of eliminated sets of hypotheses. Subjects sometimes do eliminate more than one hypothesis at a time, but considering the mental effort and memory capacity required by the normative strategy, it is not surprising that a basic +test heuristic predominates instead (Levine, 1966, 1970; Millward & Spoehr, 1973; Taplin, 1975).

The Four-Card Problem

As suggested earlier, the +test strategy applies to both Htests and Ttests. Thus, tasks that allow both are of particular interest. One example is the four-card problem (Wason, 1966, 1968; Wason & Johnson-Laird, 1972) and its descendants (e.g., Cox & Griggs, 1982; Evans & Lynch, 1973; Griggs, 1983; Griggs & Cox, 1982, 1983; Hoch & Tschirgi, 1983, 1985; Yachanin & Tweney, 1982). In these tasks, subjects are asked to determine the truth-value of the proposition "if P then Q" ($P \rightarrow Q$). For example, they may be asked to judge the truth of the following statement: "If a card has a vowel on the front, it has an even number on the back" (Wason, 1966, 1968). They are then given the opportunity to examine known cases of P, $\bar{P}$, Q, and $\bar{Q}$. For example, they can look at a card face-up with the letter E showing, face-up with the letter K, face-down with the number 4 showing, or face-down with the number 7. In our terms, this is a hypothesis-testing task in which "has an even number on the back" is the target property, and "has a vowel on the front" is the hypothesized rule that determines the target set. However, the implication $P \rightarrow Q$ is not logically equivalent to the if-and-only-if relation tested in rule discovery: P is required only to be sufficient for Q, not also necessary. Subjects nevertheless use the same basic +test approach.

From our point of view, to look at a vowel is to do a +Htest. The card with the consonant is a −Htest, the even number a +Ttest, and the odd number a −Ttest. If the +test heuristic is applied to problems of the form $P \rightarrow Q$, we would expect to find a tendency to select the +Htest and the +Ttest (P and Q), or the +Htest only (P). Indeed, these choice patterns (P and Q, or P only) are the most commonly observed in a number of replications (Evans & Lynch, 1973; Griggs & Cox, 1982; Wason, 1966, 1968; Wason & Johnson-Laird, 1972). However, there is a critical difference between the rule to be evaluated in the four-card problem and those in rule discovery. The implication $P \rightarrow Q$ is subject to only one kind of falsification, $P \cap \bar{Q}$. As a result, the +test strategy is inappropriate in this task. The only relevant tests are those that find false positives: +Htests and −Ttests (P and $\bar{Q}$, e.g., E and 7).

Earlier, we proposed that people would be able to move beyond the basic +test strategy under favorable conditions, and research on the four-card problem has demonstrated this. In particular, a number of follow-up studies have shown that a concrete context can point the way for subjects. Consider, for example, the casting of the problem at a campus pub serving beer and cola, with the proposition "if a person is drinking beer, then the person must be over 19" (Griggs & Cox, 1982). Here the real-world context alerts subjects to a critical feature of this specific task: The error of interest is "beer-drinking and not-over-19" ($P \cap \bar{Q}$). The presence of people over 19 drinking cola ($\bar{P} \cap Q$) is immaterial. In this version, people are much more likely to examine the appropriate cases, P and $\bar{Q}$ (beer drinkers and those under 19). Hoch and Tschirgi (1983, 1985) have shown similar effects for more subtle and general contextual cues as well.

Although there have been many explanations for the presence and absence of the P and Q choice pattern, a consensus seems to be emerging. The if/then construction is quite ambiguous in natural language; it often approximates a biconditional or other combination of implications (e.g., see Legrenzi, 1970; Politzer, 1986; Rumain, Connell, & Braine, 1983; Tweney & Doherty, 1983). A meaningful context disambiguates the task by indicating the practical logic of the situation. Some investigators have suggested that in an abstract or ambiguous task, people resort to a degenerate strategy of merely matching whatever is mentioned in the proposition, in other words, P and Q (Evans & Lynch, 1973; Hoch & Tschirgi, 1985; Tweney & Doherty, 1983). We suggest, however, that this heuristic of last resort is not a primitive refuge resulting from confusion or misunderstanding, but a manifestation of a more general default strategy (+testing) that turns out to be effective in many natural situations. People seem to require contextual or "extra logical" information (Hoch & Tschirgi, 1983) to help them see when this all-purpose heuristic is not appropriate to the task at hand.

Intuitive Personality Testing

Snyder, Swann, and colleagues have conducted a series of studies demonstrating that people tend to seek confirmation of

a hypothesis they hold about the personality of a target person (Snyder, 1981; Snyder & Campbell, 1980; Snyder & Swann, 1978; Swann & Giuliano, in press). For example, in some studies (Snyder, 1981; Snyder & Swann, 1978), one group of subjects was asked to judge whether another person was an extrovert, and a second group was asked to determine whether that person was an introvert. Given a list of possible interview questions, both groups tended to choose "questions that one typically asks of people already known to have the hypothesized trait" (Snyder, 1981, p. 280). For example, subjects testing the extrovert hypothesis often chose the question "What would you do if you wanted to liven things up at a party?"

This behavior is quite consistent with the +test heuristic. Someone's personality can be thought of as a set of behaviors or characteristics. To understand person A's personality is, then, to identify which characteristics in the universe of possible human characteristics belong to person A and which do not. That is, the target set (T) is the set of characteristics that are true of person A. The hypothesis "A is an extrovert" establishes a hypothesized set of characteristics (H), namely those that are true of extroverts. The goal of the hypothesis tester is, as usual, to determine if the hypothesized set coincides well with the target set. In other words, to say "A is an extrovert" is to say: "If it is characteristic of extroverts, it is likely to be true of A, and if it is not characteristic of extroverts, it is likely not true of A." Following the +test strategy, you test this by examining extrovert characteristics to see if they are true of the target person (+Htests).

The +test strategy fails in these tasks because it does not take into account an important task characteristic: Some of the available questions are nondiagnostic. The question above, for example, is not very conducive to an answer such as "Don't ask me, I never try to liven things up." Both introverts and extroverts accept the premise of the question and give similar answers (Swann, Giuliano, & Wegner, 1982). Subjects would better have chosen neutral questions (e.g., "What are your career goals?") that could be more diagnostic. However, it is not +Htesting that causes problems here; it is the mistaking of nondiagnostic questions for diagnostic ones (Fischhoff & Beyth-Marom, 1983; Swann, 1984). All the same, it is not optimal for testers to allow a general preference for +Htests to override the need for diagnostic information.

A series of recent studies suggest that, given the opportunity, people do choose to ask questions that are reasonably diagnostic; however, they still tend to choose questions for which the answer is *yes* if the hypothesized trait is correct (Skov & Sherman, 1986; Strohmer & Newman, 1983; Swann & Giuliano, in press; Trope & Bassok, 1982, 1983; Trope, Bassok, & Alon, 1984). For example, people tend to ask a hypothesized introvert questions such as "Are you shy?" Indeed, people may favor +Htesting in part because they believe +Htests to be more diagnostic in general (cf. Skov & Sherman, 1986; Swann & Giuliano, in press). Interestingly, Trope and Bassok (1983) found this +Htesting tendency only when the hypothesized traits were described as extreme (e.g., extremely polite vs. on the polite side). If an extreme personality trait implies a narrower set of behaviors and characteristics, then this is consistent with our normative analysis of +Htesting: As $p(t)$ becomes smaller, the advantage of +Htesting over −Htesting becomes greater (see Equations 1 and 2). Although only suggestive, the Trope and Bassok results may indicate that people have some salutary intuitions about how situational factors affect the +test heuristic (see also Swann & Giuliano, in press).

Learning from Outcome Feedback

So far we have only considered tasks in which the cost of information gathering and the availability of information are the same for +tests and −tests. However, several studies have looked at hypothesis testing in situations where tests are costly. Of particular ecological relevance are those tasks in which one must learn from the outcomes of one's actions. As mentioned earlier, studies by Tschirgi (1980) and Schwartz (1982) suggest that when test outcomes determine rewards as well as information, people attempt to replicate good results (reinforcement) and avoid bad results (nonreinforcement or punishment). This encourages +Htesting, because cases consistent with the best current hypothesis are believed more likely to produce the desired result.

Einhorn and Hogarth (1978; see also Einhorn, 1980) provide a good analysis of how this can lead to a conflict between two important goals: (a) acquiring useful information to revise one's hypothesis and improve long-term success, and (b) maximizing current success by acting the way you think works best. Consider the case of a university admissions panel that must select or reject candidates for admission to graduate school. Typically, they admit only those who fit their hypothesis for success in school (i.e., those who meet the selection criteria). From the point of view of hypothesis testing, the admissions panel can check on selected candidates to see if they prove worthy (+Htests). It is much more difficult to check on rejected candidates (−Htests) because they are not conveniently collected at your institution and may not care to cooperate. Furthermore, you would really have to admit them to test them, because their outcome is affected by the fact that they were rejected (Einhorn & Hogarth, 1978). In other words, −Htests would require admitting some students hypothesized to be unworthy. However, if there is any validity to the admissions committee's judgment, this would have the immediate effect of reducing the average quality of admitted students. Furthermore, it would be difficult to perform either kind of Ttest in these situations. +Ttests and −Ttests would require checking known successes and known failures, respectively, to see whether you had accepted or rejected them. As before, information about people you rejected is hard to come by and is affected by the fact that you rejected them.

The net result of these situational factors is that people are strongly encouraged to do only one kind of tests: +Htests. This limitation is deleterious to learning, because +Htests reveal only false positives, never false negatives. As in Wason's 2, 4, 6 task, this can lead to an overly restrictive rule for acceptance as you attempt to eliminate false-positive errors without knowing about the rate of false negatives.

On the other hand, our analyses suggest that there are situations in which reliance on +Htesting may not be such a serious mistake. First, it might be the case that you care more about false positives than false negatives (as suggested earlier). You may not be too troubled by the line you insert in rejection letters

stating that "Regrettably, many qualified applicants must be denied admission." In this case, +Htests are adequate because they reveal the more important errors, false positives. Even where both types of errors are important, there are many circumstances in which +Htests may be useful because false positives are more likely than false negatives (see Table 1). When $p(t) = p(h)$ and $p(t) < .5$, for example, the false-positive rate is always greater than the false-negative rate. In other words, if only a minority of applicants is capable of success in your program, and you select about the right proportion of applicants, you are more likely to be wrong about an acceptance than a rejection. As always, the effectiveness of a +test strategy depends on the nature of the task. Learning from +Htests alone is not an optimal approach, but it may often be useful given the constraints of the situation.

Judgments of Contingency

There has been considerable recent interest in how people make judgments of contingency or covariation between factors (e.g., see Alloy & Tabachnik, 1984; Arkes & Harkness, 1983; Crocker, 1981; Nisbett & Ross, 1980; Schustack & Sternberg, 1981; Shaklee & Mims, 1982), and one often-studied class of contingency tasks is readily described by the theoretical framework proposed in the present paper. These are tasks that require the subject to estimate the degree of contingency (or its presence or absence) between two dichotomous variables, on the basis of the presentation of a number of specific instances. For example, Ward and Jenkins (1965) presented subjects with the task of determining whether there was a contingency between the seeding of clouds and the occurrence of rainfall on that day. Subjects based their judgments on a series of slides, each of which indicated the state of affairs on a different day: (a) seeding + rain, (b) seeding + no rain, (c) no seeding + rain, or (d) no seeding + no rain.

In our terms, the dichotomous-contingency task can be characterized as follows: The subject is presented with a target property or event and a set of conditions that are hypothesized to distinguish occurrences of the target from nonoccurrences. In the Ward and Jenkins (1965) example, the target event is rain, and the condition of having seeded the clouds is hypothesized to distinguish rainy from nonrainy days. This task is different from rule discovery in two ways. First, the hypothesized rule is not compared to a standard of "best possible" prediction, but rather to a standard of "better than nothing." Second, the information search takes place in memory; the tester determines which information to attend to or keep track of rather than controlling its presentation. (A similar characterization is presented by Crocker, 1981.)

Despite these differences, we propose that the basic +test strategy is manifested in covariation judgment much as it is in other, more external tasks. The event types listed above can be mapped onto our division of instances into H and $\bar{H}$, T and $\bar{T}$ (see Table 3). The labels given the cells, A, B, C, and D, correspond to the terminology commonly used in studies of contingency. One possible evaluation strategy in such a problem is to think of cases in which the conditions were met (days with cloud seeding), and estimate how often those cases possessed the target property (rain). This is +Htesting: examining instances that

Table 3
The Relationship of Hypothesis-Testing Terms to Contingency Judgments

Proposed cause or condition	Target event or property: Present (T)	Target event or property: Absent ($\bar{T}$)
Present (H)	Cell A: $H \cap T$	Cell B: $H \cap \bar{T}$
Absent ($\bar{H}$)	Cell C: $\bar{H} \cap T$	Cell D: $\bar{H} \cap \bar{T}$

fit the hypothesized conditions (H: cloud seeding) to see whether they are target events (T: rain) or nontargets ($\bar{T}$: no rain). In other words, +Htesting is based on instances in cells A and B. Similarly, one could think of cases in which the target property occurred (it rained) to see whether the hypothesized conditions were met (clouds had been seeded). This is equivalent to +Ttesting, based on instances in cells A and C.

We expect, as usual, that people will favor +Htests and +Ttests over −Htests and −Ttests. We also expect that there may be a tendency toward +Htesting in particular, because of greater attention to the sufficiency of rules than to their necessity (e.g., you do not mind if it rains sometimes without seeding). Also, many contingency tasks are framed in terms of the relation between causes and effects. Htests may be more natural then, because they are consistent with the temporal order of causation, moving from known causes to possible results (cf. Tversky & Kahneman, 1980).

These hypotheses lead to some specific predictions about people's judgments of contingency. On a group level, judgments will be most influenced by the presence or absence of A-cell instances, because they are considered in both +Htests and +Ttests. B-cell and C-cell data will have somewhat less influence, because B-cell data are considered only with +Htests and C-cell only with +Ttests. If +Htests are the most popular tests, then B-cell data will receive somewhat more emphasis than C-cell data. Finally, D-cell data will have the least effect, because they are not considered in either of the favored tests. On an individual-subject level, there will be extensive use of strategies comparing cell A with cell B (+Htesting) and comparing cell A with cell C (+Ttesting).

The data from a variety of studies support these predictions. Schustack and Sternberg (1981), for example, found that the contingency judgments of subjects taken as a group were best modeled as a linear combination of the number of instances of each of the four types, with the greatest emphasis placed on A-cell, B-cell, C-cell, and D-cell data, in that order. Similar results were reported in an experiment by Arkes and Harkness (1983, Experiment 7), and in a meta-analysis of contingency–judgment tasks by Lipe (1982).

A number of studies have also examined data from individual subjects. Although some studies indicate that people are influenced almost entirely by A-cell data (Jenkins & Ward, 1965; Nisbett & Ross, 1980; Smedslund, 1963), there is now considerable evidence for the prevalence of an A − B strategy (Arkes & Harkness, 1983; Shaklee & Mims, 1981, 1982; Ward & Jenkins, 1965). This label has been applied to strategies that compare

the number of $H \cap T$ instances with the number of $H \cap \bar{T}$ (Cell A vs. Cell B) as well as strategies that compare $T \cap H$ (Cell A) with $T \cap \bar{H}$ (Cell C). The first comparison is consistent with our idea of +Htesting, the second with +Ttesting. These two kinds of comparison have not been clearly distinguished in the literature. For example, Arkes and Harkness (1983) sometimes label the condition-but-no-event cell as B, and sometimes the event-but-no-condition cell as B. However, in one study, Shaklee and Mims (1981) were able to distinguish A − B and A − C patterns in their data and found evidence of both.

Further evidence of a +test approach is found in a recent study by Doherty and Falgout (1985). They presented the Ward and Jenkins (1965) cloud-seeding task on a computer screen and enabled subjects to save instances in computer memory for later reference. Although there were large individual differences, the most common pattern was to save a record of instances in cells A and B (the results of +Htests). The second most common pattern was to save A-, B-, and C-cell instances (+Htests and +Ttests), and the third most common pattern was B and C (the falsifications from +Htests and +Ttests). Together, these 3 patterns accounted for 32 of 40 data-saving patterns in two experiments.

In contingency judgment as in rule discovery, the +test strategy can often work well as a heuristic for hypothesis testing. However, this approach can deviate appreciably from statistical standards under some circumstances. Most statistical indexes (e.g., chi-square or correlation coefficient) put equal weight on all four cells, which +testing does not. Are people capable of more sophisticated strategies? Shaklee and Mims (1981, 1982) and Arkes and Harkness (1983) describe a sum-of-diagonals strategy that generally fares well as a rough estimate of statistical contingency. However, a simple combination of +Htests and +Ttests would result in a pattern of judgments very similar to the sum-of-diagonals strategy. A stimulus set could be carefully constructed to discriminate the two, but in the absence of such studies, we suspect that many sum-of-diagonals subjects may actually be using a combination of A versus B (+Htests) and A versus C (+Ttests). This may explain why individual analyses indicate frequent use of sum-of-diagonals strategies whereas group analyses often indicate that D-cell data is given little weight. On the other hand, we would expect that subjects might use more sophisticated strategies under favorable circumstances. There is some evidence that reduced memory demands have such an effect. Contingency judgments are more sophisticated when data are presented in summary form, rather than case by case (Arkes & Harkness, 1983; Shaklee & Mims, 1981, 1982; Shaklee & Tucker, 1980; Ward & Jenkins, 1965). Also, the problem context and the wording of the question may direct attention to relevant sources of data (Arkes & Harkness, 1983; Crocker, 1982; Einhorn & Hogarth, 1986).

Further Theoretical and Empirical Questions

The concept of a general +test strategy provides an integrative interpretation for phenomena in a wide variety of hypothesis-testing tasks. This interpretation also prompts a number of new theoretical and empirical questions. There are several ways our analyses can be extended to explore further the nature of hypothesis-testing tasks and the strategies people use to accomplish them. We present a few examples here.

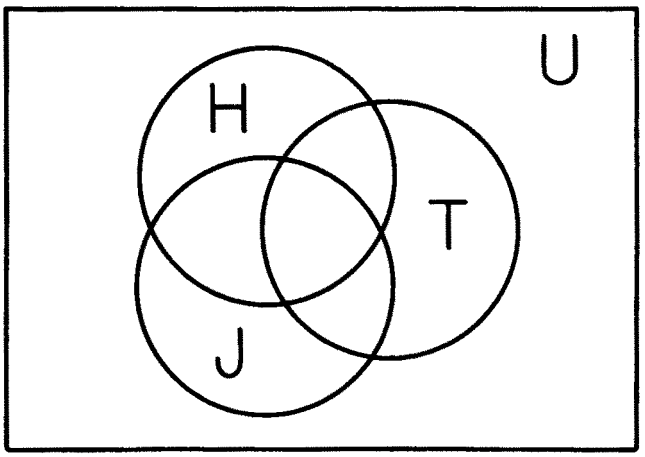


Figure 6. Representation of hypothesis testing situation involving two alternate hypotheses, R_H and R_J, specifying sets H and J, respectively.

In this article we discuss tasks in which the goal is to determine the correctness of a single hypothesis. This is a common situation, since people (including scientists) tend to view hypothesis testing in terms of verifying or falsifying one particular hypothesis (Mitroff, 1974; Tweney, 1984, 1985; Tweney & Doherty, 1983; Tweney et al., 1980). On the other hand, it would be interesting to analyze the use of simultaneous alternate hypotheses in obtaining informative tests of hypotheses (see Figure 6). The importance of specific alternatives has been emphasized in laboratory hypothesis-testing studies (e.g., Wason & Johnson-Laird, 1972, chap. 16) and in philosophical discussions (e.g., Platt, 1964). An analysis like ours could be used to examine how alternate hypotheses can increase the expected information from tests, under what circumstances an alternative is not useful (e.g., with a straw-man hypothesis), and when it would be better to simultaneously verify or falsify two alternatives rather than perform a test that favors one over the other. From a theoretical perspective, it might also be interesting to examine a situation in which a larger set of alternate hypotheses are evaluated simultaneously. This may not be representative of ordinary scientific thought, but could provide an interesting normative standard (cf. Edwards, 1965; Raiffa, 1968). It is also akin to problems commonly faced by artificial intelligence researchers in designing expert systems to perform diagnostic tasks (see, e.g., Duda & Shortliffe, 1983; Fox, 1980).

Another possible extension of these analyses is to consider standards of comparison other than "correct" or "best possible." In many situations, it may be more appropriate to ask whether or not your hypothesis is "pretty good," or "good enough," or even "better than nothing." Then, instead of comparing error rates to irreducible minima (ϵ^+ and ϵ^-), you are comparing them to other standards (s^+ and s^-). Similarly, it would be possible to consider the testing of a rule for estimating a continuous variable rather than for predicting the presence or absence of a property. What you want to know then is the expected amount of error, rather than just the probability of error.

Our theoretical analyses also suggest a number of interesting

empirical questions concerning the ways in which people adapt their strategies to the task at hand. For example, we indicate that certain task variables have a significant impact on how effective the +test strategy is in different situations. We do not know the extent to which people respond to these variables, or whether they respond appropriately. For example, do people use −Htests more when the target set is large? Will they do so if the cost of false negative guesses is made clear? Our review of existing research suggests that people may vary their approach appropriately under favorable conditions. However, there is still much to learn about how factors such as cognitive load and task-specific information affect hypothesis-testing strategies.

Finally, there is a broader context of hypothesis formation and revision that should be considered as well. We have focused on the process of finding information to test a hypothesis. The broader context also includes questions about how to interpret your findings (e.g., see Darley & Gross, 1983; Hoch & Ha, 1986; Lord et al., 1979). The astrophysicist must decide if the blur in the picture is really a planet; the interviewer must judge whether the respondent has given an extroverted answer. Moreover, questions about how hypotheses are tested are inevitably linked to questions about how hypotheses are generated. The latter sort of questions have received much less attention, however, possibly because they are harder to answer (but see, e.g., Gettys, 1983; Gettys & Fisher, 1979). Obtaining falsification is only a first step. The investigator must use that information to build a new hypothesis and must then do further testing. Thus, analyses of hypothesis testing and hypothesis generation will be mutually informative.

Conclusions

Over the past 30 years, there have been scores of studies on the nature of hypothesis testing in scientific investigation and in everyday reasoning. Many investigators talk about confirmation bias, but this term has been applied to many different phenomena in a variety of contexts. In our review of the literature, we find that different kinds of "confirmation bias" can be understood as resulting from a basic hypothesis-testing heuristic, which we call the positive test strategy. That is, people tend to test hypotheses by looking at instances where the target property is hypothesized to be present or is known to be present.

This +test strategy, in its various manifestations, has generally been regarded as incompatible with the prescription to seek disconfirmation. The central idea of this prescription is that the hypothesis tester should make a deliberate attempt to find any evidence that would falsify the current hypothesis. As we show, however, +testing does not necessarily contradict the goal of seeking falsification. Indeed, under some circumstances, +testing may be the only way to discover falsifying instances (see Figure 3). Furthermore, in probabilistic environments, it is not even necessarily the case that falsification provides more information than verification. What is best depends on the characteristics of the specific task at hand.

Our review suggests that people use the +test strategy as a general default heuristic. That is, this strategy is one that people use in the absence of specific information that identifies some tests as more relevant than others, or when the cognitive demands of the task preclude a more carefully designed strategy. Our theoretical analyses indicate that, as an all-purpose heuristic, +testing often serves the hypothesis tester well. That is probably why it persists, despite its shortcomings. For example, if the target phenomenon is relatively rare, and the hypothesis roughly matches this base rate, you are probably better off testing where you do expect the phenomenon to occur or where you know the phenomenon occurred rather than the opposite. This situation characterizes many real-world problems. Moreover, +tests may be less costly or less risky than −tests when real-world consequences are involved (Einhorn & Hogarth, 1978; Tschirgi, 1980).

Like most general-purpose heuristics, however, +testing can lead to problems when applied inappropriately. In rule discovery, it can produce misleading feedback by failing to reveal a whole class of important falsifications (violations of necessity). In propositional reasoning (e.g., the four-card problem), +testing leads to superfluous tests of necessity (+Ttests) and neglect of some relevant tests of sufficiency (−Ttests). In a variety of tasks, including concept indentification, intuitive personality testing, and contingency judgment, a +test strategy can lead to inefficiency or inaccuracy by overweighting some data and underweighting others. The consequences of using a +test strategy vary with the characteristics of the task.

Our task analyses serve two major functions. First, they highlight some of the structural similarities among diverse tasks in the broad domain of hypothesis testing. This permits integration of findings from different subareas that have so far been fairly isolated from each other. Second, our approach provides a framework for analyzing what each task requires of the subject, why people make the mistakes they do, and why changes in the structure and content of tasks sometimes produce significant changes in performance. These questions are central to understanding human hypothesis testing in the larger context of practical and scientific reasoning.

References

American Psychiatric Association (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior experience and current situational information. *Psychological Review, 91,* 112–149.

Arkes, H. R., & Harkness, A. R. (1983). Estimates of contingency between two dichotomous variables. *Journal of Experimental Psychology: General, 112,* 117–135.

Bourne, L. E., Jr. (1974). An inference model for conceptual rule learning. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 231–256). New York: Erlbaum.

Bourne, L. E., Jr., & Guy, D. E. (1968). Learning conceptual rules II: The role of positive and negative instances. *Journal of Experimental Psychology, 77,* 488–494.

Bruner, J. S. (1951). Personality dynamics and the process of perceiving. In R. R. Blake & G. V. Ramsey (Eds.), *Perception: An approach to personality* (pp. 121–147). New York: Ronald Press.

Bruner, J. S., Goodnow, J., & Austin, G. A. (1956). *A study of thinking.* New York: Wiley.

Centers for Disease Control (1986). Cases of specific notifiable diseases, United States. *Morbidity and Mortality Weekly Report, 34,* 775–777.

Cox, J. R., & Griggs, R. A. (1982). The effect of experience on performance in Wason's selection task. *Memory and Cognition, 10,* 496–502.

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

Crocker, J. (1982). Biased questions in judgment of covariation studies. *Personality and Social Psychology Bulletin, 8,* 214–220.

Darley, J. M., & Gross, P. H. (1983). A hypothesis confirming bias in labeling effects. *Journal of Personality and Social Psychology, 44,* 20–33.

Doherty, M. E., & Falgout, K. (1985, November). *Subjects' data selection strategies for assessing covariation.* Paper presented at the meeting of the Psychonomics Society, Boston, MA.

Duda, R. O., & Shortliffe, E. H. (1983). Expert systems research. *Science, 220,* 261–268.

Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processes. *Journal of Mathematical Psychology, 2,* 312–329.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representations of human judgment* (pp. 17–52). New York: Wiley.

Edwards, W., & Phillips, L. D. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72,* 346–354.

Einhorn, H. J. (1980). Learning from experience and suboptimal rules in decision making. In T. S. Wallsten (Ed.), *Cognitive processes in choice and decision behavior* (pp. 1–20). Hillsdale, NJ: Erlbaum.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review, 85,* 396–416.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99,* 3–19.

Evans, J. St. B. T., & Lynch, J. S. (1973). Matching bias in the selection task. *British Journal of Psychology, 64,* 391–397.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90,* 239–260.

Fox, J. (1980). Making decisions under the influence of memory. *Psychological Review, 87,* 190–211.

Gettys, C. F. (1983). *Research and theory on predecisional processes* (Rep. No. TR-11-30-83). Norman: University of Oklahoma, Decision Processes Laboratory.

Gettys, C. F., & Fisher, S. D. (1979). Hypothesis generation and plausibility assessment. *Organizational Behavior and Human Performance, 24,* 93–110.

Gorman, M. E. (1986). How the possibility of error affects falsification on a task that models scientific problem-solving. *British Journal of Psychology, 77,* 85–96.

Gorman, M. E., & Gorman, M. E. (1984). A comparison of disconfirmatory, confirmatory, and a control strategy on Wason's 2-4-6 task. *Quarterly Journal of Experimental Psychology, 36A,* 629–648.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Griggs, R. A. (1983). The role of problem content in the selection task and the THOG problem. In J. St. B. T. Evans (Ed.), *Thinking and reasoning: Psychological approaches* (pp. 16–43). London: Rutledge & Kegan Paul.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology, 73,* 407–420.

Griggs, R. A., & Cox, J. R. (1983). The effect of problem content on strategies in Wason's selection task. *Quarterly Journal of Experimental Psychology, 35,* 519–533.

Hardin, C. L. (1980). Rationality and disconfirmation. *Social Studies of Science, 10,* 509–514.

Hoch, S. J., & Ha, Y.-W. (1986). Consumer learning: Advertising and the ambiguity of product experience. *Journal of Consumer Research, 13,* 221–233.

Hoch, S. J., & Tschirgi, J. E. (1983). Cue redundancy and extra logical inference in a deductive reasoning task. *Memory & Cognition, 11,* 200–209.

Hoch, S. J., & Tschirgi, J. E. (1985). Logical knowledge and cue redundancy in deductive reasoning. *Memory & Cognition, 13,* 453–462.

Hogarth, R. M. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin, 90,* 197–217.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied, 79* (1, Whole No. 594).

Kern, L. H. (1983, November). *The effect of data error in inducing confirmatory inference strategies in scientific hypothesis testing.* Paper presented at the meeting of the Society for the Social Studies of Science, Blacksburg, VA.

Klayman, J. (1986). *An information-theory analysis of the value of information in hypothesis testing* (Working Paper No. 119a). Chicago, IL: University of Chicago, Graduate School of Business, Center for Decision Research.

Klayman, J., & Ha, Y.-W. (1985, August). *Strategy and structure in rule discovery.* Paper presented at the Tenth Research Conference on Subjective Probability, Utility and Decision Making, Helsinki, Finland.

Lakatos, I. (1970). Falsification and methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of scientific knowledge* (pp. 91–196). New York: Cambridge University Press.

Legrenzi, P. (1970). Relations between language and reasoning about deductive rules. In G. B. Flores d'Arcais & W. J. M. Levelt (Eds.), *Advances in psycholinguistics* (pp. 322–333). Amsterdam: North Holland.

Levine, M. (1966). Hypothesis behavior by humans during discrimination learning. *Journal of Experimental Psychology, 71,* 331–338.

Levine, M. (1970). Human discrimination learning: The subset-sampling assumption. *Psychological Bulletin, 74,* 397–404.

Lipe, M. G. (1982). *A cross-study analysis of covariation judgments* (Working Paper No. 96). Chicago, IL: University of Chicago, Graduate School of Business, Center for Decision Research.

Lord, C., Ross, L., & Lepper, M. (1979). Biased assimilation and attitude polarization: The effect of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37,* 2098–2109.

Lucas, T., & Whittemore, H. (1985). *Tornado!* (NOVA program No. 1217). Boston: WGBH Transcripts.

Mahoney, M. J. (1976). *Scientist as subject: The psychological imperative.* Cambridge, MA: Ballinger.

Mahoney, M. J. (1979). Psychology of the scientist: An evaluative review. *Social Studies of Science, 9,* 349–375.

Mahoney, M. J. (1980). Rationality and authority: On the confusion of justification and permission. *Social Studies of Science, 10,* 515–518.

Millward, R. B., & Spoehr, K. T. (1973). The direct measurement of hypothesis-testing strategies. *Cognitive Psychology, 4,* 1–38.

Mitroff, I. (1974). *The subjective side of science.* Amsterdam: Elsevier.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology, 29,* 85–95.

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

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

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

Politzer, G. (1986). Laws of language use and formal logic. *Journal of Psycholinguistic Research, 15,* 47–92.

Popper, K. R. (1959). *The logic of scientific discovery.* New York: Basic Books.

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

Raiffa, H. (1968). *Decision analysis.* Reading, MA: Addison-Wesley.

Ross, L., & Lepper, M. R. (1980). The perseverance of beliefs: Empirical and normative considerations. In R. A. Shweder (Ed.), *Fallible judgment in behavioral research: New directions for methodology of social and behavioral science* (Vol. 4, pp. 17–36). San Francisco: Jossey-Bass.

Rumain, B., Connell, J., & Braine, M. D. S. (1983). Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults: *If* is not the biconditional. *Developmental Psychology, 19,* 471–481.

Sagan, C. (1980). *Cosmos.* New York: Random House.

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General, 110,* 101–120.

Schwartz, B. (1981). Control of complex, sequential operants by systematic visual information in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes, 7,* 31–44.

Schwartz, B. (1982). Reinforcement-induced behavioral stereotypy: How not to teach people to discover rules. *Journal of Experimental Psychology: General, 111,* 23–59.

Shaklee, H., & Mims, M. (1981). Development of rule use in judgments of covariation between events. *Child Development, 52,* 317–325.

Shaklee, H., & Mims, M. (1982). Sources of error in judging event covariations: Effects of memory demands. *Journal of Experimental Psychology: Learning, Memory & Cognition, 8,* 208–224.

Shaklee, H., & Tucker, D. (1980). A rule analysis of judgments of covariation between events. *Memory & Cognition, 8,* 459–467.

Simon, H. A. (1973). Does scientific discovery have a logic? *Philosophy of Science, 40,* 471–480.

Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis confirmatory strategies and perceived hypothesis confirmation. *Journal of Experimental Social Psychology, 22,* 93–121.

Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology, 4,* 165–173.

Snyder, M. (1981). Seek and ye shall find: Testing hypotheses about other people. In E. T. Higgins, C. P. Heiman, & M. P. Zanna (Eds.), *Social cognition: The Ontario symposium on personality and social psychology* (pp. 277–303). Hillsdale, NJ: Erlbaum.

Snyder, M., & Campbell, B. H. (1980). Testing hypotheses about other people: The role of the hypothesis. *Personality and Social Psychology Bulletin, 6,* 421–426.

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

Strohmer, D. C., & Newman, L. J. (1983). Counselor hypothesis-testing strategies. *Journal of Counseling Psychology, 30,* 557–565.

Swann, W. B., Jr. (1984). Quest for accuracy in person perception: A matter of pragmatics. *Psychological Review, 91,* 457–477.

Swann, W. B., Jr., & Giuliano, T. (in press). Confirmatory search strategies in social interaction: How, when, why and with what consequences. *Journal of Social and Clinical Psychology.*

Swann, W. B., Jr., Giuliano, T., & Wegner, D. M. (1982). Where leading questions can lead: The power of conjecture in social interaction. *Journal of Personality and Social Psychology, 42,* 1025–1035.

Taplin, J. E. (1975). Evaluation of hypotheses in concept identification. *Memory & Cognition, 3,* 85–96.

Trabasso, T., & Bower, G. H. (1968). *Attention in learning.* New York: Wiley.

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

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

Trope, Y., Bassok, M., & Alon, E. (1984). The questions lay interviewers ask. *Journal of Personality, 52,* 90–106.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development, 51,* 1–10.

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

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (Vol. 1, pp. 49–72). Hillsdale, NJ: Erlbaum.

Tweney, R. D. (1984). Cognitive psychology and the history of science: A new look at Michael Faraday. In H. Rappard, W. van Hoorn, & S. Bem (Eds.), *Studies in the history of psychology and the social sciences* (pp. 235–246). The Hague: Mouton.

Tweney, R. D. (1985). Faraday's discovery of induction: A cognitive approach. In D. Gooding & F. James (Eds.), *Faraday rediscovered* (pp. 159–209). London: MacMillan.

Tweney, R. D., & Doherty, M. E. (1983). Rationality and the psychology of inference. *Synthese, 57,* 139–161.

Tweney, R. D., Doherty, M. E., & Mynatt, C. R. (1982). Rationality and disconfirmation: Further evidence. *Social Studies of Science, 12,* 435–441.

Tweney, R. D., Doherty, M. E., Worner, W. J., Pliske, D. B., Mynatt, C. R., Gross, K. A., & Arkkelin, D. L. (1980). Strategies of rule discovery in an inference task. *Quarterly Journal of Experimental Psychology, 32,* 109–123.

Vogel, R., & Annau, Z. (1973). An operant discrimination task allowing variability of response patterning. *Journal of the Experimental Analysis of Behavior, 20,* 1–6.

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology, 19,* 231–241.

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

Wason, P. C. (1962). Reply to Wetherick. *Quarterly Journal of Experimental Psychology, 14,* 250.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135–151). Harmondsworth, Middlesex, England: Penguin.

Wason, P. C. (1968). On the failure to eliminate hypotheses—A second look. In P. C. Wason & P. N. Johnson-Laird (Eds.), *Thinking and reasoning* (pp. 165–174). Harmondsworth, Middlesex, England: Penguin.

Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content.* London: Batsford.

Wetherick, N. E. (1962). Eliminative and enumerative behavior in a conceptual task. *Quarterly Journal of Experimental Psychology, 14,* 246–249.

Yachanin, S. A., & Tweney, R. D. (1982). The effect of thematic content on cognitive strategies in the four-card selection task. *Bulletin of the Psychonomic Society, 19,* 87–90.

(Appendix follows on next page)

Appendix

Two Measures of the Expected Impact of a Test

Assume that you have a hypothesized rule, R_H, and some subjective degree of belief that this rule is the best possible, $p(R_H = R_B)$. Your goal is to achieve the maximum degree of certainty that $R_H = R_B$ or $R_H \neq R_B$. Suppose that you perform a +Htest, and receive a falsification (Fn, +H). Then, according to Bayes's equation, your new degree of belief should be

$$p(R_H = R_B|Fn, +H) = \frac{p(Fn, +H|R_H = R_B)}{p(Fn, +H)} \cdot p(R_H = R_B). \quad (A1)$$

According to earlier definitions, $p(Fn, +H|R_H = R_B) = p(\bar{t}|b) = \epsilon^+$, and $p(Fn, +H) = p(\bar{t}|h) = z^+$. Thus

$$p(R_H = R_B|Fn, +H) = \frac{\epsilon^+}{z^+} \cdot p(R_H = R_B). \quad (A2)$$

Similarly, if your +Htest yields verification,

$$p(R_H = R_B|Vn, +H) = \frac{1 - \epsilon^+}{1 - z^+} \cdot p(R_H = R_B). \quad (A3)$$

By definition, $\epsilon^+ \leq z^+$, so verifications produce an increased degree of belief that $R_H = R_B$ (or no change) and falsification a decrease in belief (or no change). For −Htests, revisions are equivalent but depend on ϵ^- and z^- rather than ϵ^+ and z^+.

Using the expected change in belief (EΔP) as a measure of informativeness (as defined in the text),

$$\Delta P_{Fn,+H} = \left| p(R_H = R_B) - \frac{\epsilon^+}{z^+} p(R_H = R_B) \right|$$

$$= p(R_H = R_B) \cdot \left[1 - \frac{\epsilon^+}{z^+} \right],$$

$$\Delta P_{Vn,+H} = \left| p(R_H = R_B) - \frac{1 - \epsilon^+}{1 - z^+} \cdot p(R_H = R_B) \right|$$

$$= p(R_H = R_B) \cdot \left[\frac{1 - \epsilon^+}{1 - z^+} - 1 \right], \text{ and}$$

$$E\Delta P_{+H} = p(Fn|+H) \cdot \Delta P_{Fn,+H} + p(Vn|+H) \cdot \Delta P_{Vn,+H}$$

$$= z^+ \left[1 - \frac{\epsilon^+}{z^+} \right] \cdot p(R_H = R_B)$$

$$+ (1 - z^+) \left[\frac{1 - \epsilon^+}{1 - z^+} - 1 \right] \cdot p(R_H = R_B)$$

$$= (z^+ - \epsilon^+) \cdot p(R_H = R_B) + (z^+ - \epsilon^+) \cdot p(R_H = R_B)$$

$$= p(R_H = R_B) \cdot 2(z^+ - \epsilon^+). \quad (A4)$$

Similarly,

$$E\Delta P_{-H} = p(R_H = R_B) \cdot 2(z^- - \epsilon^-). \quad (A5)$$

An alternate measure of impact, diagnosticity, is frequently used in Bayesian analyses. An alternate form of Bayes's theorem states that

$$\frac{p(R_H = R_B|Result)}{p(R_H \neq R_B|Result)} = \frac{p(Result|R_H = R_B)}{p(Result|R_H \neq R_B)} \cdot \frac{p(R_H = R_B)}{p(R_H \neq R_B)} \quad (A6)$$

$$\Omega' = \quad LR \quad \cdot \Omega$$

The likelihood ratio (LR) is the basis of the diagnosticity measure. It is equal to the ratio of revised odds (Ω') to prior odds (Ω). A likelihood ratio of 1 means the result has no impact on your beliefs; it is nondiagnostic. The further from 1 the likelihood ratio is, the greater the event's impact.

Edwards (1968; Edwards & Phillips, 1966) suggests that subjective uncertainty may be better represented by log odds than by probabilities or raw odds, based on evidence that subjective estimates made on such a scale tend to conform better to normative specifications. Following this suggestion, diagnosticity can be measured as the magnitude of the change in log-odds (ΔL) that an event would engender, which is equivalent to the magnitude of the log likelihood ratio, |log LR|. If, for instance, you performed a +Htest and received falsification, the diagnosticity of this datum would be

$$\Delta L_{Fn,+H} = \log \frac{p(R_H = R_B)}{1 - p(R_H = R_B)} - \log \frac{p(R_H = R_B|Fn, +H)}{1 - p(R_H = R_B|Fn, +H)}. \quad (A7)$$

For ease of exposition, we will use the letter C to stand for the subjective probability $p(R_H = R_B)$. Following equations A2 and A3 above,

$$\Delta L_{Fn,+H} = \log \frac{C}{1 - C} - \log \frac{\epsilon^+/z^+ \cdot C}{1 - (\epsilon^+/z^+ \cdot C)} \quad (A8)$$

and

$$\Delta L_{Vn,+H} = \log \frac{\frac{1 - \epsilon^+}{1 - z^+} \cdot C}{1 - \left(\frac{1 - \epsilon^+}{1 - z^+} \cdot C \right)} - \log \frac{C}{1 - C} \quad (A9)$$

Parallel to our earlier analyses, we can define the expected change in log-odds (EΔL) for a +Htest as $p(Fn|+Htest) \cdot \Delta L_{Fn,+H} + p(Vn|+Htest) \cdot \Delta L_{Vn,+H}$. That is,

$$E\Delta L_{+H} = z^+ \Delta L_{Fn,+H} + (1 - z^+) \Delta L_{Vn,+H}. \quad (A10)$$

Accordingly, the expected change in log-odds for −Htests can be calculated by substituting ϵ^- for ϵ^+ and z^- for z^+ in Equations A8, A9, and A10.

EΔL increases monotonically with increasing z, except for some small, local violations when C is very low, z is very high, and ϵ is near .5 (rather degraded conditions). EΔL decreases monotonically with increasing ϵ. Thus, as in earlier analyses, more information is expected from the test with the higher z and the lower ϵ. The exact trade-off between z and ϵ is complex, however. Under most circumstances, the component due to falsifications ($z^+\Delta L_{Fn,+H}$ for +Htests or $z^-\Delta L_{Fn,-H}$ for −Htests) is greater than the component due to verification [$(1 - z^+)\Delta L_{Vn,+H}$ or $(1 - z^-)\Delta L_{Vn,-H}$, respectively]. That is, more information is expected to come from falsification, overall, than from verification with this measure.

Received January 25, 1986
Revision received August 6, 1986 ■