Psychological Review
2003, Vol. 110, No. 3, 526–535

0033-295X/03/$12.00 DOI: 10.1037/0033-295X.110.3.526

Hypothesis Testing and Theory Evaluation at the Boundaries: Surprising Insights From Bayes's Theorem

David Trafimow
New Mexico State University

Because the probability of obtaining an experimental finding given that the null hypothesis is true [$p(F\backslash H_0)$] is not the same as the probability that the null hypothesis is true given a finding [$p(H_0\backslash F)$], calculating the former probability does not justify conclusions about the latter one. As the standard null-hypothesis significance-testing procedure does just that, it is logically invalid (J. Cohen, 1994). Theoretically, Bayes's theorem yields $p(H_0\backslash F)$, but in practice, researchers rarely know the correct values for 2 of the variables in the theorem. Nevertheless, by considering a wide range of possible values for the unknown variables, it is possible to calculate a range of theoretical values for $p(H_0\backslash F)$ and to draw conclusions about both hypothesis testing and theory evaluation.

Despite a variety of different criticisms, the standard null-hypothesis significance-testing procedure (NHSTP) has dominated psychology over the latter half of the past century. Although NHSTP has its defenders when used "properly" (e.g., Abelson, 1997; Chow, 1998; Hagen, 1997; Mulaik, Raju, & Harshman, 1997), it has also been subjected to virulent attacks (Bakan, 1966; Cohen, 1994; Rozeboom, 1960; Schmidt, 1996). For example, Schmidt and Hunter (1997) argue that NHSTP is "logically indefensible and retards the research enterprise by making it difficult to develop cumulative knowledge" (p. 38). According to Rozeboom (1997), "Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students" (p. 336). The most important reason for these criticisms is that although one can calculate the probability of obtaining a finding given that the null hypothesis is true, this is not equivalent to calculating the probability that the null hypothesis is true given that one has obtained a finding. Thus, researchers are in the position of rejecting the null hypothesis even though they do not know its probability of being true (Cohen, 1994). One way around this problem is to use Bayes's theorem to calculate the probability of the null hypothesis given that one has obtained a finding, but using Bayes's theorem carries with it some problems of its own, including a lack of information necessary to make full use of the theorem. Nevertheless, by treating the unknown values as variables, it is possible to conduct some analyses that produce some interesting conclusions regarding NHSTP. These analyses clarify the relations between NHSTP and Bayesian theory and quantify exactly why the standard practice of rejecting the null hypothesis is, at times, a highly questionable procedure. In addition, some surprising findings come out of the analyses that bear on issues pertaining not only to hypothesis testing but also to the amount of information gained from findings and theory evaluation. It is possible that the implications of the following analyses for information gain and theory evaluation are as important as the NHSTP debate.

I thank Susan Schibel for her valuable help in preparing the figures. I also thank Doug Gillan, Victor Johnston, Tim Ketelaar, Laura Madson, and Jim McDonald for their helpful comments. Finally, I thank a reviewer for suggestions regarding the title.

Correspondence concerning this article should be addressed to David Trafimow, Department of Psychology, New Mexico State University, MSC 3452, P.O. Box 30001, Las Cruces, New Mexico 88003-8001. E-mail: trafimow@crl.nmsu.edu

Hypothesis Testing

The first section of this article concerns two questions pertaining to hypothesis testing. First, when is it acceptable and when is it not acceptable to reject the null hypothesis? Second, how much information does one gain when conducting an experiment?

Rejecting the Null Hypothesis

At a bare minimum, NHSTP includes the following steps:

1. Propose a hypothesis to be (hopefully) supported.
2. Propose a null hypothesis (H_0) to be (hopefully) rejected (the hypothesis and H_0 are supposed to be defined such that they are mutually exclusive and exhaustive).
3. Collect the data.
4. Compute the probability of obtaining the finding (e.g., a difference between the experimental and control condition) given that H_0 is true [$p(F\backslash H_0)$].
5. If $p(F\backslash H_0) < .05$, reject H_0 and conclude that the alternative hypothesis (H_1) is true.

As it has been outlined above, NHSTP may seem eminently logical. Obviously, if H_0 has been eliminated [because $p(F\backslash H_0) < .05$], then the remaining hypothesis is supported. Unfortunately, matters are not so simple. If it were impossible to obtain the finding given the null hypothesis [$p(F\backslash H_0) = 0$], then the logic of NHSTP would be unassailable and provable by the following reasoning:

1. If H_0 is true, the finding cannot happen. (Premise 1)

2. The finding happens. (Premise 2)
3. Therefore H_0 is not true. (Conclusion 1)
4. Either H_0 or H_1 must be true. (by definition)
5. H_0 is not true. (from 3)
6. Therefore H_1 must be true. (Conclusion 2)

Unfortunately, when dealing with probabilities rather than with certainties, the above reasoning does not hold. If $p(F\backslash H_0) = .05$, $p(H_0\backslash F)$ can take on any value between 0 and 1, depending on two other probabilities: the prior probability of the null hypothesis [$p(H_0)$] and the probability of obtaining the finding if H_0 is not true [$p(F\backslash -H_0)$]. To see why this is so, consider Bayes's theorem below:

$$p(H_0\backslash F) = p(F\backslash H_0)p(H_0)/[p(F\backslash H_0)p(H_0)+p(F\backslash -H_0)p(-H_0)]. \quad (1)$$

Note that by definition, $p(-H_0)$ = probability of the alternative hypothesis [$p(H_1)$] = $1 - p(H_0)$, so that Equation 2 is also true:

$$p(H_0\backslash F) = p(F\backslash H_0)p(H_0)/[p(F\backslash H_0)p(H_0) + p(F\backslash -H_0)(1 - p(H_0))]. \quad (2)$$

In summary, Equation 2 implies that one needs to know three probabilities to calculate $p(H_0\backslash F)$. These are $p(F\backslash H_0)$, $p(H_0)$, and $p(F\backslash -H_0)$. The first of these probabilities can be calculated from the data with *t* tests, *F* ratios, and the like, but one rarely has enough information to calculate $p(H_0)$ or $p(F\backslash -H_0)$. Thus, it is understandable that NHSTP rather than Bayesian analysis is the order of the day—after all, researchers can perform the analyses.

Unfortunately, the fact that NHSTP is easy does not mean that it is valid. Even if $p(F\backslash H_0)$ is .05, it is easy to choose values for $p(H_0)$ or $p(F\backslash -H_0)$ that will result in $p(H_0\backslash F)$ being quite a bit larger than .05, and in an extreme enough case, this value can even equal 1. To see this quickly, suppose either $p(H_0)$ is 1 or $p(F\backslash -H_0)$ is 0. In the first case, $1 - p(H_0) = 1 - 1 = 0$, and so the whole term containing that value (to the right of the plus sign in the denominator of Equation 2) equals 0. Thus, Equation 2 reduces to $p(F\backslash H_0)\ p(H_0)/p(F\backslash H_0)\ p(H_0) = 1$. In the second case, in which $p(F\backslash -H_0) = 0$, the whole term containing that value again equals 0, and so Equation 2 again reduces to $p(F\backslash H_0)\ p(H_0)/p(F\backslash H_0)\ p(H_0) = 1$. Thus, in either case, even if $p(F\backslash H_0) < .05$, $p(H_0\backslash F)$ equals 1. Clearly, rejecting the null hypothesis [because $p(F\backslash H_0) < .05$] when it is certainly true is a bad idea.

An astute reader might object that substituting 1 for $p(H_0)$ or 0 for $p(F\backslash -H_0)$ is extreme and that NHSTP might fare better with less extreme values. In fact, as I shall demonstrate, NHSTP does fare better with less extreme values instantiated into Equation 2. Figure 1 shows all of the values $p(H_0\backslash F)$ can have assuming that $p(F\backslash H_0)$ is set at .05 (the conventional alpha level); that $p(F\backslash -H_0)$ equals .1 (top curve), .2 (next to the top curve), . . . , .9 (bottom curve); and that $p(H_0)$ is greater than or equal to 0 and less than or equal to 1. The horizontal axis represents the values that $p(H_0)$ can have and the vertical axis represents the probabilities that $p(H_0\backslash F)$ can have. Finally, the top [$p(F\backslash -H_0) = .1$] and bottom [$p(F\backslash -H_0) = .9$] curves are labeled, but there was no room to label the other curves [$p(F\backslash -H_0) = .2, \ldots, .8$], and so they are not labeled.

To begin with, Figure 1 shows two basic tendencies. First, as $p(H_0)$ increases, so does $p(H_0\backslash F)$. Second, as $p(F\backslash -H_0)$ decreases, $p(H_0\backslash F)$ increases. So it is immediately clear that large values for $p(H_0)$ or small values for $p(F\backslash -H_0)$ imply large values for $p(H_0\backslash F)$. There is also a joint effect of $p(H_0)$ and $p(F\backslash -H_0)$ on $p(H_0\backslash F)$. As Figure 1 indicates, when $p(F\backslash -H_0)$ is low (e.g., .1), most values for $p(H_0)$ result in values for $p(H_0\backslash F)$ being greater than .05. For example, any value of .1 or higher for $p(H_0)$ will result in a final value for $p(H_0\backslash F)$ being greater than .05. Thus, looking at the top curve implies that NHSTP usually results in the rejection of the null hypothesis when it is quite likely to be true.

On the other hand, a look at the bottom curve in Figure 1, in which $p(F\backslash -H_0) = .9$, suggests a more optimistic evaluation of NHSTP. Values for $p(H_0)$ can go almost up to .5 before $p(H_0\backslash F)$ begins to exceed .05. Thus, in this case, almost half of the values $p(H_0)$ can take on result in a conservative rejection rate when NHSTP is used. Further, high values for $p(H_0\backslash F)$ do not occur until $p(H_0)$ exceeds .8 [at this value, $p(H_0\backslash F) = .18$], and extremely high values for $p(H_0\backslash F)$ do not occur until $p(H_0)$ exceeds .9 [at this value, $p(H_0\backslash F) = .33$].

Is NHSTP too liberal? Figure 1 shows that this depends on $p(H_0)$ and $p(F\backslash -H_0)$. NHSTP can be extremely liberal or extremely conservative depending on these values. So NHSTP is too liberal, too conservative, or just right, depending on what one believes these values are, in the kinds of studies and experiments that get published in psychology journals.

Information Gain

It is widely accepted that scientists conduct studies or experiments to gain information; but what do we as scientists mean when we talk about information gain? One way of defining information gain is in terms of the change in the probability of H_0 from before the finding was obtained [the prior probability of H_0 or $p(H_0)$] to its probability after the finding is obtained [the posterior probability of H_0 or $p(H_0\backslash F)$]. If the posterior probability is very different from the prior probability, then a lot of information from the experiment has been gained. A more precise way of saying this is that obtaining the finding that H_1 predicts (and that H_0 does not predict) provides maximal information when the posterior probability of H_0 is much less than the prior probability of H_0 [i.e., when $p(H_0) - p(H_0\backslash F)$ = high number]. Equation 3 summarizes these points:

$$\text{Information gain (I)} = p(H_0) - p(H_0\backslash F). \quad (3)$$

Equation 3 can also be expressed in terms of the same variables in Equation 2. So substituting everything to the right of the equal sign in Equation 2 into Equation 3 yields Equation 4:

$$I = p(H_0) - \{p(F\backslash H_0)p(H_0)/[p(F\backslash H_0)p(H_0) + p(F\backslash -H_0)(1 - p(H_0))]\}. \quad (4)$$

Equations 3 and 4 define information gain in terms of change from the prior to the posterior probability of H_0. It is also possible to define information gain in terms of the change in prior to posterior probability of H_1 by substituting H_1 for H_0 in Equation 4, which leads to Equation 5:

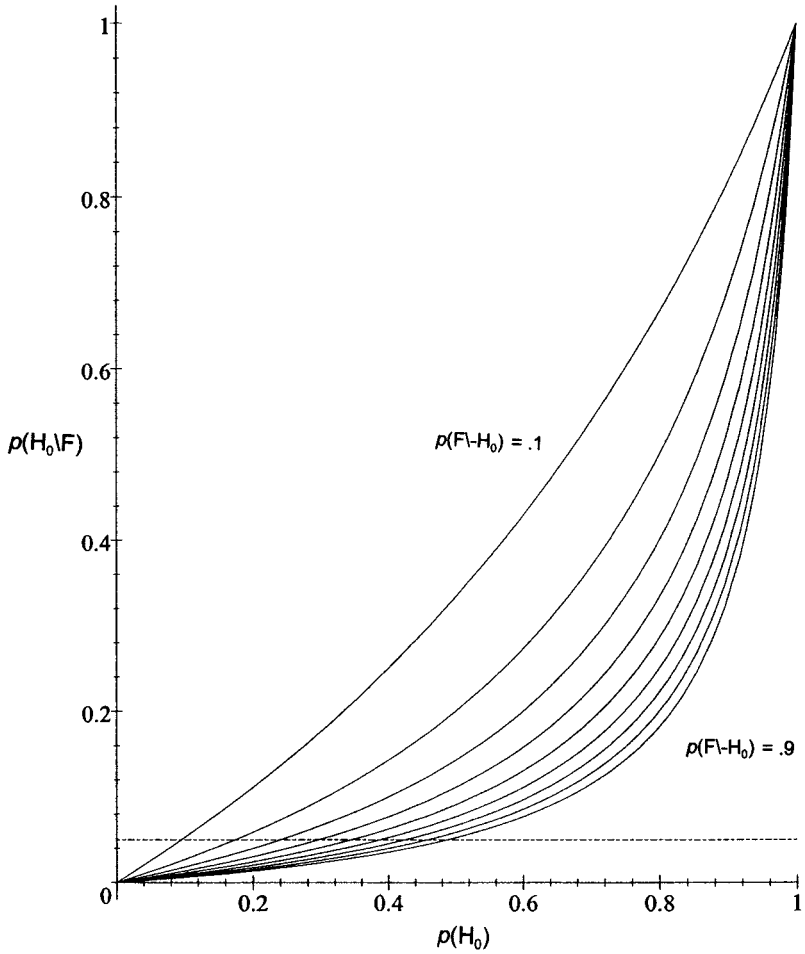

Figure 1. Curves representing nine equations, based on Equation 2, for calculating $p(H_0\backslash F)$ as a function of $p(H_0)$, $p(F\backslash H_0)$, and $p(F\backslash -H_0)$. These variables took on the following values: $p(H_0)$ varied between 0 and 1; $p(F\backslash H_0)$ was set at the conventional alpha level of .05; and $p(F\backslash -H_0)$ was set at .1 (top curve), .2 (next to top curve), . . . , .9 (bottom curve). The dashed line represents $p(H_0\backslash F) = .05$. $p(H_0\backslash F)$ = the probability that the null hypothesis is true given a finding; $p(H_0)$ = prior probability of the null hypothesis; $p(F\backslash H_0)$ = probability of the finding given the null hypothesis; $p(F\backslash -H_0)$ = probability of obtaining the finding given that the null hypothesis is not true.

$$I = p(H_1) - \{p(F\backslash H_1)p(H_1)/[p(F\backslash H_1)p(H_1) + p(F\backslash -H_1)(1 - (H_1))]\}. \quad (5)$$

However, even if one wishes to think of information gain in terms of H_1, it is still useful to be able to use the same variables that were used in Figure 1. So, $1 - p(H_0)$ can be substituted for $p(H_1)$, $p(F\backslash -H_0)$ for $p(F\backslash H_1)$, $p(F\backslash H_0)$ for $p(F\backslash -H_1)$, and $p(H_0)$ for $1 - p(H_1)$, which leads to Equation 6:

$$I = [1 - p(H_0)] - \{p(F\backslash -H_0)(1 - p(H_0))/[p(F\backslash -H_0)(1 - p(H_0)) + p(F\backslash H_0)p(H_0)]\}. \quad (6)$$

Note that expressing information gain in terms of H_1 rather than H_0, as in Equation 6, renders the change in probability of H_1 as a negative rather than a positive number. This happens because successfully obtaining a finding increases the posterior probability of H_1 over its prior probability unless the prior probability is 0 or 1.

Figure 2 illustrates information gain in terms of H_1 but still uses the variables from Figure 1. Thus, the following were instantiated into Equation 6 and are represented in Figure 2: $p(F\backslash H_0)$ was kept at .05, $p(H_0)$ was allowed to vary from 0 to 1, and $p(F\backslash -H_0)$ was set at .1 (top curve), .2 (next to top curve), . . . , .9 (bottom curve). As in Figure 1, the top $[p(F\backslash -H_0)]$ and bottom $[p(F\backslash -H_0)]$ curves are labeled, and the other curves are not.

To understand the implications of Figure 2, consider what happens to information gain as $p(F\backslash -H_0)$ increases and $p(H_0)$ increases. As $p(F\backslash -H_0)$ increases, information gain also increases. Note that the bottommost curve represents the greatest information gain. This makes good intuitive sense; as the likelihood of the finding given that H_0 is wrong (and therefore that H_1 is right) increases, the finding provides a much stronger case for H_1.

Now consider information gain as $p(H_0)$ increases. Figure 2 shows that when $p(H_0)$ is a small number, information gain also tends to be small. But when $p(H_0)$ gets closer to 1 (but not too

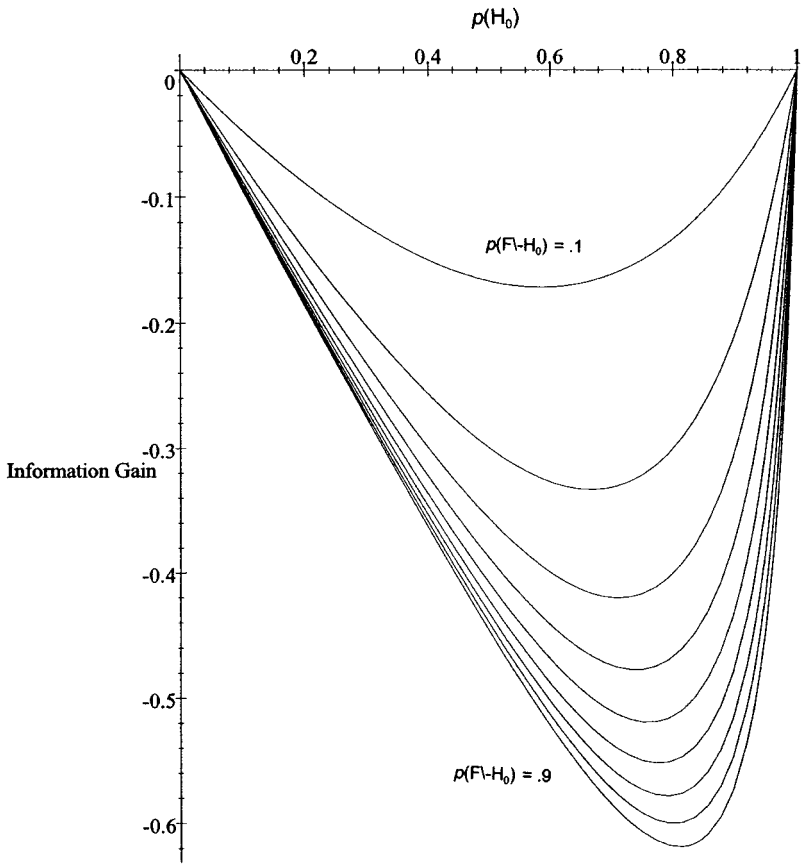

Figure 2. Curves representing nine equations, based on Equation 6, for calculating information gain as a function of $p(H_0)$, $p(F\backslash H_0)$, and $p(F\backslash{-}H_0)$. These variables took on the following values: $p(H_0)$ varied between 0 and 1; $p(F\backslash H_0)$ was set at the conventional alpha level of .05; and $p(F\backslash{-}H_0)$ was set at .1 (top curve), .2 (next to top curve), . . . , .9 (bottom curve). $p(H_0)$ = prior probability of the null hypothesis; $p(F\backslash H_0)$ = probability of the finding given the null hypothesis; $p(F\backslash{-}H_0)$ = probability of obtaining the finding given that the null hypothesis is not true.

close), information gain is maximized. This is particularly true when $p(F\backslash{-}H_0)$ is a large number. [It is important to note that the "best" value for $p(H_0)$ for maximizing information gain is dependent on $p(F\backslash{-}H_0)$. It should also be noted that the best value would change if $p(F\backslash H_0)$ were set at values different from .05.] In terms of H_1, what this means is that when researchers propose "obvious" hypotheses (i.e., the prior probability of H_0 is low and that of H_1 is high), they fail to gain much information even when they obtain the desired finding. However when researchers propose nonobvious hypotheses so that the prior probability of the null hypothesis is high (but not too high) and that of the alternative hypothesis is low, then obtaining the desired finding results in a large amount of information gain. The philosophical moral of Figure 2 is that researchers should try to test nonobvious hypotheses.[1]

Theory Evaluation

The foregoing discussion focused on the relations between hypotheses and findings but ignored theories. However, the reason researchers propose hypotheses and perform experiments is to evaluate their theories (at least this is supposed to be so in basic psychology journals). So suppose that a researcher has derived a hypothesis from a theory and wishes to know the extent to which the theory is supported by the hypothesis. The posterior probability of the theory, given that the hypothesis is true, can be expressed by Bayes's theorem as shown in Equation 7. In essence, Equation 7 is similar to Equation 2 except that the idea is to find the probability of the theory, given that the derived hypothesis is true $[p(T\backslash H_1)]$, in terms of the prior probability of the theory $[p(T)]$, the probability of the derived hypothesis given the theory $[p(H_1\backslash T)]$, and the probability of the derived hypothesis if the theory is not true $[p(H_1\backslash{-}T)]$:

$$p(T\backslash H_1) = p(H_1\backslash T)p(T)/[p(H_1\backslash T)p(T) + p(H_1\backslash{-}T)(1 - p(T))]. \quad (7)$$

[1] It should be noted, however, that this recommendation involves a good deal of risk to the researcher. The testing of nonobvious hypotheses entails a substantial probability that H_0 will fail to be rejected. It is often difficult to publish failed attempts to reject H_0, and so the researcher who adopts the recommended strategy should be prepared to deal with this consequence.

Equation 7 allows one to perform analyses that are somewhat analogous to those performed in the previous section, but with regard to theories and hypotheses rather than hypotheses and findings. In the following subsection, I explore the posterior probability of the theory being true given that the hypothesis is true. In the subsection after that, I explore the change in probability of the theory after the hypothesis derived from the theory is supported.

The Posterior Probability of the Theory

It is probably a myth in science that research hypotheses are arrived at solely by deduction from theories (Faust, 1984). If this myth were true, it would imply that given the truth of a theory, the hypothesis must be true. Accept the myth, for a moment, and assume that $p(H_1\backslash T) = 1$. In that case, Equation 7 reduces to Equation 8:

$$p(T\backslash H_1) = p(T)/[p(T) + p(H_1\backslash -T)(1 - p(T))]. \quad (8)$$

On the basis of Equation 8, Figure 3A presents the posterior probability of the theory [$p(T\backslash H_1)$] as a function of the prior probability of the theory [$p(T)$, which is allowed to vary from 0 to 1] and $p(H_1\backslash -T)$, which is set at .1 (top curve), .2 (next to top curve), . . . , .9 (bottom curve). The horizontal axis represents the prior probability of the theory, and the vertical axis represents the theory's posterior probability. As in Figures 1 and 2, the top [$p(H_1\backslash -T) = .1$] and bottom [$p(H_1\backslash -T) = .9$] curves are labeled, and the other curves are not.

Not surprisingly, as the prior probability of the theory increases, so does the posterior probability. It is more interesting to note that the posterior probability of the theory is strongly dependent on the probability of the hypothesis when the theory is not true [$p(H_1\backslash -T)$]. As $p(H_1\backslash -T)$ decreases, $p(T\backslash H_1)$ increases. Thus, Figure 3A demonstrates the importance of testing theories with hypotheses that are unlikely to be true if the theory is not true.

Of course, Figure 3A depends on the unlikely assumption that $p(H_1\backslash T) = 1$. As several philosophers and psychologists have demonstrated, hypotheses are derived from a combination of a theory and "auxiliary" assumptions outside the theory that estab-

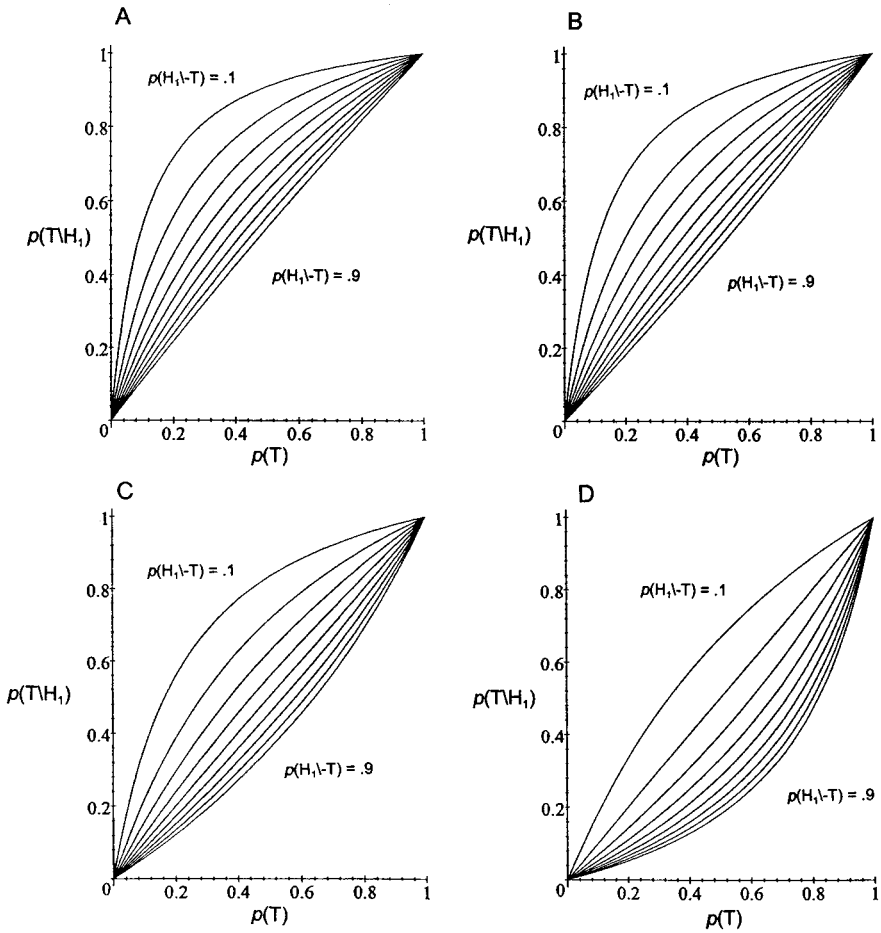

Figure 3. These four panels were based on Equation 7 in which $p(H_1\backslash T)$ was set at 1 (A), .8 (B), .5 (C), and .2 (D). Within each of these panels, the curves represent nine equations for calculating $p(T\backslash H_1)$ as a function of $p(T)$, which ranges between 0 and 1, and $p(H_1\backslash -T)$, which equals .1 (top curve), .2 (next to top curve), . . . , .9 (bottom curve). $p(H_1\backslash T)$ = probability of the alternative hypothesis given the theory; $p(T\backslash H_1)$ = the probability of the theory given the alternative hypothesis; $p(T)$ = prior probability of the theory; $p(H_1\backslash -T)$ = probability of the alternative hypothesis given that the theory is not true.

lish initial conditions, assumptions about the measurement of variables, and others (Hempel, 1965; Lakatos, 1978; Meehl, 1990, 1997; Popper, 1959, 1962). Meehl (1997) has provided a list of these auxiliary assumptions. Unless one is willing to assume that the probability of the truth of these auxiliary assumptions equals 1, it is necessary to assume that $p(H_1\backslash T)$ is less than 1. So, Figure 3B assumes that $p(H_1\backslash T) = .8$, Figure 3C assumes that $p(H_1\backslash T) = .5$, and Figure 3D assumes that $p(H_1\backslash T) = .2$. The points made above with regard to Figure 3A also come across in Figures 3B, 3C, and 3D. It is interesting to note, however, that as $p(H_1\backslash T)$ decreases (across Figures 3A to 3D), the posterior probability of the theory also decreases. There is an interesting interaction whereby the shape of the curves depends on both $p(H_1\backslash T)$ and $p(H_1\backslash -T)$. When $p(H_1\backslash T)$ is high and $p(H_1\backslash -T)$ is low, then the curves tend to rise quickly before flattening out. This means that even when the prior probability of the theory is low, the posterior probability may nevertheless be quite high. Conversely, when $p(H_1\backslash T)$ is low and $p(H_1\backslash -T)$ is high, the curves are rather flat to begin with before rising substantially. This means that even with a respectable prior probability of the theory, the posterior probability of the theory may nevertheless not be very high. Finally, a low $p(H_1\backslash T)$ can be counterbalanced by an even lower $p(H_1\backslash -T)$, and a high $p(H_1\backslash -T)$ can be counterbalanced by an even higher $p(H_1\backslash T)$, with all of these effects depending on the prior probability of the theory.

Change in Probability of the Theory as a Result of the Hypothesis

By reasoning similar to that in the previous section, it is possible to derive an equation that represents the change from the prior probability of the theory to its posterior probability (given that the hypothesis has been shown to be true). This difference is $p(T\backslash H_1) - p(T)$ and represents the gain in confidence one can have in the theory as a result of obtaining findings that support the veracity of the hypothesis (C = change in confidence). This reasoning is summarized in Equation 9:

$$C = p(T\backslash H_1) - p(T). \quad (9)$$

Substituting the right half of Equation 7 for $p(T\backslash H_1)$ renders Equation 10:

$$C = \{p(H_1\backslash T)p(T)/[p(H_1\backslash T)p(T) + p(H_1\backslash -T)(1 - p(T))]\} - p(T). \quad (10)$$

Analogous to Figures 3A, 3B, 3C, and 3D in which $p(H_1\backslash T)$ was 1, .8, .5, .2, respectively, Figures 4A, 4B, 4C, and 4D represent change in confidence as a function of each of these values. Within each figure, $p(T)$ was allowed to vary from 0 to 1, and $p(H_1\backslash -T)$ was set at .1 (top curve), .2 (next to top curve), . . . , .9 (bottom curve). Thus, the horizontal axis represents the prior probability of the theory [$p(T)$], and the vertical axis represents the change in confidence in the theory (C). Similar to the other figures, the top [$p(H_1\backslash -T) = .1$] and bottom [$p(H_1\backslash -T) = .9$] curves are labeled, and the other curves are not.

First, consider Figure 4A. This figure demonstrates two important points. First, as the probability of the hypothesis when the theory is not true decreases [$p(H_1\backslash -T)$ is a low number], one can have much more confidence in the theory. Thus, Figure 4A demonstrates the importance of proposing H_1s that are unlikely to be true if the theory is not true. Second, note that as the prior probability of the theory itself decreases (as long as it doesn't get too close to 0), the amount of change in confidence is maximized. However, Figures 4B, 4C, and 4D demonstrate that this effect is more complicated than might be inferred from Figure 4A.

Now consider Figure 4B, in which $p(H_1\backslash T)$ was set at .8. As was true with Figure 4A, the greatest amount of positive change in confidence in the theory is engendered when the probability of the hypothesis if the theory is not true is minimized. This conclusion holds for all of the Figures (i.e., Figures 4A to 4D). Note also that, in general, the changes in confidence in the theory are lower in Figure 4B than in Figure 4A. This means that as the hypothesis is less tightly derived from the theory [$p(H_1\backslash T)$ is a lower number], support for the hypothesis provides decreased support for the theory. Although this is not surprising, another conclusion implied by Figure 4B is quite surprising. Specifically, when $p(H_1\backslash -T)$ is set at .9 (the lowest curve), the change in confidence in the theory actually dips into negative territory. This means that if the hypothesis derived from the theory is shown to be true, it actually decreases one's confidence in the theory. How can this happen if the hypothesis was actually derived from the theory, and in a reasonably tight way, too [remember that $p(H_1\backslash T)$ was set at .8 for Figure 4B]? The answer lies in the probability of the hypothesis given that the theory is not true, which is .9 for the bottom curve in Figure 4B. What this means is that the hypothesis can be more tightly derived if the theory is not true than if the theory is true, and so support for the hypothesis actually militates against the theory.

Figures 4C and 4D make this point clear in a more dramatic way: Several of the curves dip into negative territory, and if $p(H_1\backslash -T)$ is set at .9, these dips imply an impressive degree of negative change in confidence in the theory. The philosophical moral of Figures 4A to 4D is that not only are hypotheses that are unlikely to be true if the theory is not true required to substantially increase one's confidence in the theory but also failure to propose such hypotheses can actually decrease the posterior probability of the theory.

There is an additional, although less important, conclusion that can be drawn from Figures 4A to 4D. Specifically, as one progresses from Figure 4A to Figure 4D, the point of maximum change in confidence occurs farther along the horizontal axis. Thus, to obtain the maximum change in confidence in a theory, the best prior value for the theory changes depending on how tightly the hypothesis is derived from the theory.

Discussion

The foregoing analyses imply a variety of conclusions relating to NHSTP, information gain, theory evaluation, and change in confidence in theories. These will be addressed in turn.

1. There has been a great deal of debate over the desirability of NHSTP (Abelson, 1997; Chow, 1998; Cohen, 1994; Mulaik et al., 1997; Rozeboom, 1997; Schmidt, 1996; Schmidt & Hunter, 1997; Wilkinson & The Task Force on Statistical Inference, 1999). The analyses reported here, particularly in Figure 1, not only cast doubt on the validity of NHSTP but also specify the amount by which researchers are wrong when they reject the null hypothesis under various values for $p(H_0)$ and $p(F\backslash -H_0)$ even when the finding is significant at $\alpha = .05$. Whether the "wrongness" is in a liberal or

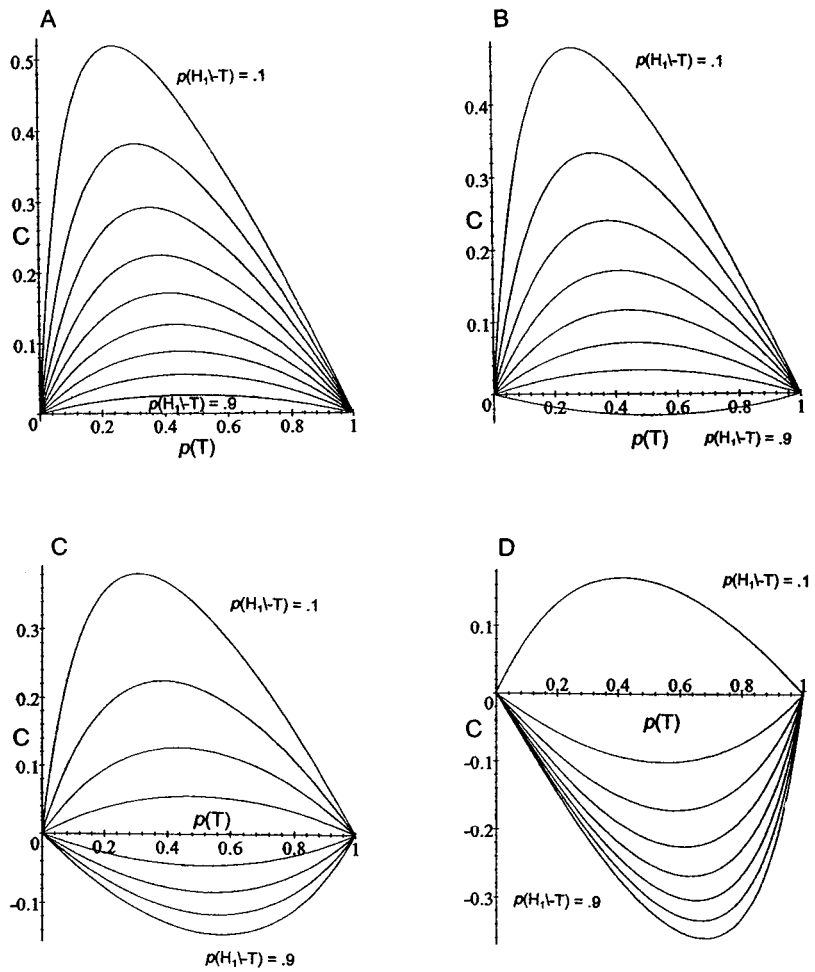

Figure 4. These four panels were based on Equation 10 in which $p(H_1\backslash T)$ was set at 1 (A), .8 (B), .5 (C), and .2 (D). Within each of these panels, curves represent nine equations for calculating change in confidence (C) as a function of $p(T)$, which ranges between 0 and 1, and $p(H_1\backslash -T)$, which equals .1 (top curve), .2 (next to top curve), . . . , .9 (bottom curve). $p(H_1\backslash T)$ = probability of the alternative hypothesis given the theory; $p(T)$ = prior probability of the theory; $p(H_1\backslash -T)$ = probability of the alternative hypothesis given that the theory is not true.

conservative direction depends, of course, on the values $p(H_0)$ and $p(F\backslash -H_0)$ are assumed to have.

It is important to make clear, however, that the problems with rejecting H_0 that are engendered by NHSTP do not necessarily constitute an argument against the common practice of calculating p values [$p(F\backslash H_0)$]. To see this, consider the reasons why researchers calculate p. Two reasons are usually given, and I will show that one of them is more appropriate than the other. The first use of p, and the one that seems to be emphasized in most statistics textbooks, has to do with drawing conclusions about populations. H_0 and H_1 are both statements about populations, and so using p to reject one in favor of the other is equivalent to drawing a conclusion about populations. As I have shown, this use of p is questionable at best.

The second reason researchers compute p is that they often conduct experiments to support a theory, and therefore they wish to argue that the truth of the theory is what is ultimately responsible for their finding. To make this argument more compelling, researchers must rule out as many alternative explanations for the finding as possible. One alternative explanation that must be rendered less plausible is that the finding is due to chance. If p is small, then it is less plausible that the finding is due to chance, and so other explanations are more plausible (e.g., the theory or alternative theories). This seems like an appropriate use of p, and I am not advocating that researchers stop calculating this value. In summary, although it seems reasonable to calculate p for the purpose of rendering chance as a less plausible explanation for findings, it seems less reasonable to perform the full NHSTP that ultimately results in a conclusion about populations. Possibly, however, in those cases in which there is good reason to believe that $p(H_0)$ is low and $p(F\backslash -H_0)$ is high, NHSTP might be reasonable (but see Point 2 below).

2. Figure 2 demonstrates that greater information gain is engendered when $p(F\backslash -H_0)$ is maximized. Figure 2 also suggests that as

the prior probability of the null hypothesis [$p(H_0)$] gets closer to 1 (but not too close), information gain is maximized. This latter fact points to a surprising dilemma for those who wish to perform NHSTP. That is, to justify NHSTP, Figure 1 shows that it is necessary to have a low value for $p(H_0)$. However, doing this results in very little information gain. Thus, the valid performance of NHSTP implies little information gain, and gaining a lot of information implies an invalid use of NHSTP. There seem to be two solutions to this dilemma. The most obvious solution is to not use NHSTP. However, for those who still wish to perform NHSTP (despite Figure 1), note that as $p(\text{F}\backslash{-}H_0)$ increases, both $p(H_0\backslash\text{F})$ decreases, and information gain increases. So with a large enough value for $p(\text{F}\backslash{-}H_0)$, perhaps it is possible to validly perform NHSTP and nevertheless gain a respectable amount of information. One way of increasing $p(\text{F}\backslash{-}H_0)$ is to drastically increase the number of participants. To see why this might work, consider that as the sample sizes approach the population sizes, $p(\text{F}\backslash{-}H_0)$ approaches 1. (Remember that this assumes H_0 is not true. If H_0 is true, then increasing sample sizes will tend to reinforce the lack of a difference.) Of course, this brings up other issues that will not be discussed here (e.g., How meaningful is an effect if it takes a large number of participants to get it? How many participants are needed? How would researchers know how many participants are needed given that they usually do not have the required values to obtain this number?).

In addition to sample sizes, population variances affect $p(\text{F}\backslash{-}H_0)$. To see this quickly, imagine that Population A has a mean of 10, and Population B has a mean of 12. In addition, imagine that there is 0 variance in both populations. In this case, even with sample sizes of only 1, it would be guaranteed that the difference between the two sample means would equal 2 (this is because with 0 variance, all of the scores from A would be 10 and from B would be 12). To state this principle in more general terms, as the population variances approach 0, $p(\text{F}\backslash{-}H_0)$ approaches 1 (when the populations really are different). I hasten to add, however, that it may require unusual cleverness on the part of the researcher to design experiments with low population variances (a high degree of control over irrelevant variables could be a big help here). Furthermore, in the case of nonexperimental designs, control over population variances is particularly unlikely to be under the researcher's control.

How quickly can the positive effects of increasing sample sizes and decreasing population variances on $p(\text{F}\backslash{-}H_0)$ be realized? This depends on two other variables. The first of these is the actual difference between the two populations. The second is the size of the difference between samples that the researcher needs to qualify as being extreme enough to qualify as *F* (which, among other things, depends partly on what alpha is set at). As the actual population difference increases or the required difference between samples decreases, $p(\text{F}\backslash{-}H_0)$ increases, and fewer participants and/or greater population variances become more tolerable. A mathematical analysis of how sample sizes, population variances, population differences, and required differences between samples vary to affect $p(\text{F}\backslash{-}H_0)$ is beyond the scope of this article. It is sufficient for now to merely note that all of these factors affect $p(\text{F}\backslash{-}H_0)$; consequently, all of them affect both $p(H_0\backslash\text{F})$ and information gain, and therefore all of them provide possible ways of dealing with the dilemma.

3. There is an issue that often comes up about whether obtaining $p < .05$ with a small number of participants is as "good" as obtaining it with a large number of participants. Points 1 and 2 suggest that the answer depends on whether one is talking about rejecting H_0 or rendering chance as a less probable explanation of the data. In the former case, the number of participants affects $p(\text{F}\backslash{-}H_0)$, and consequently it also affects $p(H_0\backslash\text{F})$. Therefore, the same *p* values, when obtained with different numbers of participants, can have quite different implications for NHSTP. In the latter case, however, $p = p$ regardless of the number of participants used to obtain *p*. (And some would even argue that obtaining a small *p* value with only a few participants implies an increased effect size.) Given the foregoing argument that rejecting the null hypothesis is rarely justified anyway, the implication is that in most studies, a small *p* value obtained with a small sample size is as good as a small *p* value obtained with a large sample size.

4. For conducting basic research, hypothesis testing derives much of its importance from its connection to a more general theory. For example, a researcher probably does not care whether participants recall more words in an experimental condition than in a control condition. The reason such a finding may matter is because of the implications this difference might have for a theory about how information is encoded and recalled. Thus, it is of crucial importance to consider the relation between theories and the hypotheses that are derived from them. Figures 3A to 3D demonstrate a number of important issues that were mentioned previously, although I will only discuss two here. First and most important, the probability of the finding given that the theory is not true has an extremely strong effect on the posterior probability of the theory. One reason this is important is that researchers tend to focus most of their attention on the probability of the hypothesis given that the theory is true, when the probability of the hypothesis when the theory is not true is just as important. Put another way, it is absolutely crucial for researchers to derive hypotheses that are likely to not be true, absent the theory, if they want to have a chance at providing an impressive argument for that theory. Although Lakatos (1978), Meehl (1990, 1997), and Platt (1964) have argued for such strong hypotheses (I am using the word *strong* to designate hypotheses that are unlikely to be true if the theory is wrong), it is not always easy to find research articles in which this has been done. In much published research, the hypotheses would be likely to be true even if the theories were not true. Roberts and Pashler (2000) have argued convincingly that this is particularly a problem in research evaluating quantitative theories with free parameters, which can be made to fit any plausible set of findings. According to these researchers, "The need to make predictions that are at least a little implausible seems to have been overlooked by quantitative theorists" (p. 360). However, as Meehl (1997) suggested, the problem is not limited to only this domain: "Most of psychology is nowhere near that 'ideal Popperian' stage of theory testing yet" (p. 415; also see Gergen, 1978; Gigerenzer, 1998; Schaller & Crandall, 1998; L. Wallach & Wallach, 2001; M. A. Wallach & Wallach, 1998). The upshot is that supporting the truth of hypotheses in psychology has generally provided only weak support for the theories from which they were derived. (Given the cognitive limits of researchers reviewed by Faust, 1984, this is likely to be a problem in many sciences, so I do not want to imply that psychologists are worse at research than are scientists in other domains.) Possibly one reason for the lack of emphasis on strong

hypotheses is that nobody has ever quantified the consequences of failing to derive strong hypotheses from theories. (I will admit, however, that the difficulty in calculating the probability of the hypothesis absent the theory might also be a reason.) I hope that the present quantitative demonstration will have an effect that the qualitative arguments of the past have not.

A second but related issue pertains to the relation between $p(H_1\backslash T)$ and $p(H_1\backslash -T)$. Even if $p(H_1\backslash T)$ is low, showing that the hypothesis is true might still result in a reasonable posterior probability of the theory if $p(H_1\backslash -T)$ is substantially lower than $p(H_1\backslash T)$. On the other hand, even if $p(H_1\backslash T)$ is high, if $p(H_1\backslash -T)$ is also high, then the posterior probability of the theory may still remain low. All of these issues, of course, are also dependent on the prior probability of the theory. These issues become even clearer when framed in terms of change in confidence in the theory, which will be discussed next.

5. It is not only the case, as Figures 3A to 3D demonstrate, that weak hypotheses [in which $p(H_1\backslash -T)$ is not a low number] fail to result in impressive posterior probabilities of the theories (unless of course, the prior probability of the theory was high to begin with), but Figures 4A to 4D demonstrate that weak hypotheses also fail to change the level of confidence one can have in theories. Worse yet, if a hypothesis is sufficiently weak [$p(H_1\backslash -T)$ is a high number], support for the hypothesis derived from a theory can actually decrease the confidence that can be placed on that theory. Thus, whether one thinks in terms of the posterior probability of the theory or the difference between the posterior probability of the theory and the prior probability of the theory, the message is the same—researchers should propose hypotheses that are unlikely to be true if the theory is not true. Similarly, the ratio of $p(H_1\backslash T)$ to $p(H_1\backslash -T)$ is quite important. If the ratio is high (much greater than 1), then change in confidence in the theory is likely to be impressively positive. If that ratio is low (much less than 1), then change in confidence in the theory is likely to be negative—the opposite of what the researcher would have liked to achieve.

This discussion brings up an interesting way that theories can be tested against each other. It is best, of course, to propose a hypothesis that is tightly derived from one theory (T1), and that should not be true if an alternative theory (T2) is true. However, even if the researcher cannot think of a way to do this, there may be an easier way of nevertheless providing an argument that T1 is better than T2. That is, if the hypothesis is more likely under T1 than T2, even if it is somewhat likely under T2, then $p(H_1\backslash T1) > p(H_1\backslash T2)$, and so confirming the hypothesis would support T1 over T2. How much would T1 be supported over T2? That would depend on the judged ratio of $p(H_1\backslash T1)/p(H_1\backslash T2)$. Also, the importance of this ratio itself depends on the prior probabilities of T1 and T2.

6. Before I conclude, the foregoing analyses imply a strong distinction between research that is (a) not theoretical or exploratory versus research that is (b) not theoretical but is exploratory versus research that is (c) theoretical. Suppose researchers employed by a drug company want to show that a particular drug cures a particular disease, but they are not concerned with why this happens. Thus, they are not concerned with information gain or theory testing. They just want to have a high posterior probability that H_1 is true (and H_0 is false). Figure 1 implies that the researchers will be in best shape if the prior probability of H_0 is low (so the prior probability of H_1 is high). However, Figure 2 shows that this will result in little information gain. If one is performing exploratory research and wishes to maximize information gain, Figure 2 demonstrates that it is better to have H_0s that have a high prior probability (so the prior probability of H_1 is low). For example, if our hypothetical researchers wanted to maximize information gain, they would be better off testing a drug that had not been tested before rather than testing a drug for which there was already good prior evidence of effectiveness.

Finally, if one wishes to test theories, it is best to derive hypotheses that are likely to be true if the theory is true but that are likely to be false if the theory is false. Consider the periodic table published by Mendeléev in 1869. Mendeléev assumed that the crucial property that distinguished the elements from each other was not atomic weight, as had been previously assumed, but rather valence (the tendency of elements to combine with other elements). To make the elements in each column of his table equivalent with regard to valence, he was forced to put an element of greater atomic weight ahead of one of lesser atomic weight (e.g., tellurium was put ahead of the lighter iodine to keep tellurium in the valence = 2 column and iodine in the valence = 1 column). Worse yet, he found it necessary to leave gaps in the table. Instead of apologizing for the gaps, however, Mendeléev boldly asserted that they represented elements that had not yet been discovered. On the basis of the position of the gaps in the table, he went on to predict exactly the characteristics each element would have when discovered (e.g., melting point, boiling point, valence, pattern of spectral lines in emissions when heated). Clearly, these predictions were extremely likely to be wrong if the theory was wrong. Although the world of chemistry remained skeptical of both the theory and the predictions, within a few years new elements were discovered, and their characteristics conformed to Mendeléev's bold predictions in every way. After that, nobody could doubt the utility of Mendeléev's periodic table. It is probably worth commenting that the low prior probability of the theory (at least in the judgment of other chemists) was doubtless an aid to Mendeléev in making such strong predictions.

In summary, the distinctions between research that is (a) not theoretical or exploratory versus research that is (b) not theoretical but is exploratory versus research that is (c) theoretical suggest that it is often a mistake to try to make the same hypothesis serve different purposes. This is not to prevent, however, a researcher from proposing different hypotheses within the same project for different purposes. Table 1 presents a summary of desiderata for each type of research.

Conclusion

The fact that Bayes's theorem provides the foundation for all of the foregoing analyses and conclusions should not be interpreted to mean that Bayesian statistical analyses should be routinely used in the science of psychology. My best guess is that some of the necessary information, particularly $p(H_0)$ and $p(F\backslash -H_0)$ or numbers from which these probabilities can be estimated, are often lacking, and consequently a Bayesian approach cannot be used. I do claim, however, that Bayesian thinking can be valuable even if the analyses cannot be carried out for the particular experiment at hand. For example, the Bayesian analyses presented earlier not only suggest possible problems with NHSTP but also demonstrate when these potential problems become actual problems and when

Table 1
Desiderata for Values of $p(H_0)$, $p(H_1)$, $p(F\backslash H_0)$, $p(F\backslash H_1)$, $p(H_1\backslash F)$, I, $p(H_1\backslash T)$, $p(H_1\backslash -T)$, p(T), and C, Depending on Whether the Research Is Not Exploratory or Theoretical (X), Exploratory but Not Theoretical (Y), or Theoretical (Z)

	Research type		
Value	X	Y	Z
$p(H_0)$	Low	High[a]	High[a]
$p(H_1)$	High	Low[b]	Low[b]
$p(F\backslash H_0)$	Low	Low	Low
$p(F\backslash H_1)$	High	High	High
$p(H_1\backslash F)$[c]	High	High	High
I[c]	—	High	High
$p(H_1\backslash T)$	—	—	High
$p(H_1\backslash -T)$	—	—	Low
p(T)	—	—	Low[b]
C[c]	—	—	High

Note. $p(H_0)$ = prior probability of the null hypothesis; $p(H_1)$ = prior probability of the alternative hypothesis; $p(F\backslash H_0)$ = probability of the finding given the null hypothesis; $p(F\backslash H_1)$ = probability of the finding given the alternative hypothesis; $p(H_1\backslash F)$ = probability of the alternative hypothesis given the finding; I = information gain; $p(H_1\backslash T)$ = probability of the alternative hypothesis given the theory; $p(H_1\backslash -T)$ = probability of the alternative hypothesis given that the theory is not true; p(T) = prior probability of the theory; C = change in confidence in the theory. Dashes indicate that the variables to which the values pertain are not important for the type of research. [a] The value should be close to 1, but not too close (see Figure 2). [b] The value should be close to 0, but not too close (see Figure 2 and Figures 4A–4D). [c] $p(H_1\backslash F)$ is the goal and most important value for Research Type X; I is the goal and most important value for Research Type Y; C is the goal for Research Type Z.

they do not. Furthermore, these analyses quantify the size of the problems and suggest possible solutions. However, the implications of Bayesian thinking are not limited to NHSTP. Rather, a variety of conclusions pertaining to hypotheses testing, information gain, and theory evaluation also came out of thinking issues through in a Bayesian way. An example of a contribution of Bayesian thinking to theory testing was that the importance of proposing strong hypotheses became clear. One might well conclude that if one believes that there is a high proportion of weak hypotheses in even the top journals in psychology, then weak hypotheses are as much of a problem as NHSTP.

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117–141). Mahwah, NJ: Erlbaum.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437.

Chow, S. L. (1998). Précis of statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences, 21,* 169–239.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Faust, D. (1984). *The limits of scientific reasoning.* Minneapolis: University of Minnesota Press.

Gergen, K. J. (1978). Toward generative theory. *Journal of Personality and Social Psychology, 36,* 1344–1360.

Gigerenzer, G. (1998). Surrogates for theories. *Theory and Psychology, 8,* 195–204.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

Hempel, C. G. (1965). *Aspects of scientific explanation, and other essays in the philosophy of science.* New York: Free Press.

Lakatos, I. (1978). *The methodology of scientific research programmes.* Cambridge, England: Cambridge University Press.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry, 1,* 108–141.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–425). Mahwah, NJ: Erlbaum.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65–115). Mahwah, NJ: Erlbaum.

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

Popper, K. R. (1959). *The logic of scientific discovery.* New York: Basic Books. (Original work published 1934)

Popper, K. R. (1962). *Conjectures and refutations.* New York: Basic Books.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107,* 358–367.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335–391). Mahwah, NJ: Erlbaum.

Schaller, M., & Crandall, C. S. (1998). On the purposes served by psychological research and its critics. *Theory and Psychology, 8,* 205–212.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1,* 115–129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight objections to the discontinuation of significance testing in the analysis of research data. In L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: Erlbaum.

Wallach, L., & Wallach, M. A. (2001). Experiments in social psychology: Science or self-deception? *Theory and Psychology, 11,* 451–473.

Wallach, M. A., & Wallach, L. (1998). When experiments serve little purposes: Misguided research in mainstream psychology. *Theory and Psychology, 8,* 183–194.

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Received October 8, 2001
Revision received July 14, 2002
Accepted July 29, 2002 ■