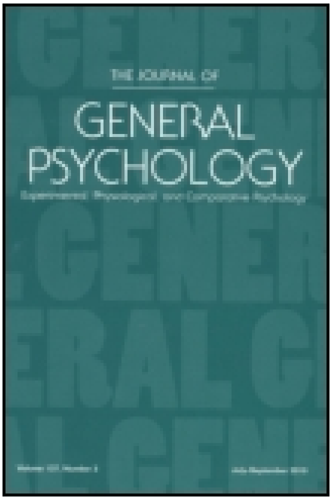

The Journal of General Psychology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/vgen20

A Test of the Null Hypothesis Significance Testing Procedure Correlation Argument

David Trafimow [a] & Stephen Rice [a]
[a] New Mexico State University
Published online: 07 Aug 2010.

The Journal of General Psychology, 2009, *136*(3), 261–269

A Test of the Null Hypothesis Significance Testing Procedure Correlation Argument

DAVID TRAFIMOW
STEPHEN RICE
New Mexico State University

ABSTRACT. Some supporters of the null hypothesis significance testing procedure recognize that the logic on which it depends is invalid because it only produces the probability of the data if given the null hypothesis and not the probability of the null hypothesis if given the data (e.g., J. Krueger, 2001). However, the supporters argue that the procedure is good enough because they believe that the probability of the data if given the null hypothesis correlates with the probability of the null hypothesis if given the data. The present authors' main goal was to test the size of the alleged correlation. To date, no other researchers have done so. The present findings indicate that the correlation is unimpressive and fails to provide a compelling justification for computing *p* values. Furthermore, as the significance rule becomes more stringent (e.g., .01, .001), the correlation decreases.

Keywords: Bayes, p_{rep}, *p* value

THERE HAS BEEN A GREAT DEAL OF CONTROVERSY about the null hypothesis significance testing procedure (NHSTP) involving a large number of supporters (e.g., Abelson, 1997; Chow, 1998; Hagen, 1997; Mulaik, Raju, & Harshman, 1997) and a large number of detractors (e.g., Bakan, 1966; Cohen, 1994; Rozeboom, 1960; Schmidt, 1996; Schmidt & Hunter, 1997). Stated briefly, NHSTP requires that the researcher propose a null hypothesis and an alternative hypothesis, collect data, and use the data to compute the probability of obtaining a finding as extreme or more extreme than the one actually obtained, given that the null hypothesis is true. If this probability is low (e.g., $p < .05$), then the researcher rejects the null hypothesis in favor of the alternative hypothesis. Otherwise, the null hypothesis is not rejected.

Address correspondence to David Trafimow, Department of Psychology, MSC 3452, New Mexico State University, PO Box 30001, Las Cruces, NM 88003-8001, USA; trafimow@nmsu.edu (e-mail).

Possibly, the most compelling argument against NHSTP is that it is logically invalid. Stated simply, the fact that a rare finding, given the null hypothesis, has been obtained does not justify the conclusion that the null hypothesis is likely to be false. This has been pointed out numerous times (for a review, see Nickerson, 2000), and Trafimow (2003) even provided a quantitative demonstration of the invalidity.

For those unfamiliar with the details of the controversy concerning NHSTP, it may be useful to consider an example. Suppose that a researcher randomly assigns participants to receive money or not to receive money and then measures how much the participants like the experiment. The experimental hypothesis is that participants will like the experiment better if they get money than if they do not. The null hypothesis is that the money has no effect (or more technically, that the two groups are drawn from the same population). The researcher performs a statistical significance test and finds that p value is .05, which meets the usual criterion for statistical significance. Can the researcher then conclude that the probability of the null hypothesis being true, given this finding, is .05 (or some other very small number)? It would be nice if this were so because researchers would then have a good reason to reject the null hypothesis.

However, this conclusion is not justified. To see why, it is useful to consider exactly what p is: the probability of the finding (or a more extreme finding) given that the null hypothesis is true. Or, in terms of the present example, it is the probability of obtaining the finding given that the samples of participants who were or were not given money are from the same population. Contrary to common belief, p is not the probability that the samples are from the same population given the finding that was obtained. Or, to state the problem another way, the probability of the finding given the null hypothesis—p or $p(F|H_0)$—is not the same thing as the probability of the null hypothesis given the finding—$p(H_0|F)$. Consequently, a low value for p, which is the same thing as $p(F|H_0)$, does not allow the researcher to validly conclude that $p(H_0|F)$ also has a low value, and therefore the rejection of the null hypothesis cannot be justified on this basis. The only way to traverse the distance from the former value (which can be obtained from an experiment) to the latter (which cannot be obtained from an experiment) is by using Bayes's theorem. The details of that theorem are not essential at present, except that its use requires the researcher to have a value for the prior probability of the null hypothesis—a value that is generally unknowable (for a review, see Trafimow, 2006).

Many of the researchers who support NHSTP fully understand that it is logically invalid for the reason previously presented but suggest that the logical invalidity need not cause the researcher to completely abandon NHSTP. For example, Krueger (2001) pointed out correctly that NHSTP results in p values—or $p(F|H_0)$—that do not allow researchers to validly draw conclusions about $p(H_0|F)$, which is what researchers really need to know to reject the null hypothesis. But Krueger also suggested correctly that these two values are correlated. Because $p(F|H_0)$ is correlated with $p(H_0|F)$, and because the data necessary to actually obtain $p(H_0|F)$ are generally unobtainable, it may be reasonable for researchers to settle for $p(F|H_0)$.

Let us reiterate the correlation argument in the specific context of the example. There is wide agreement that the obtained p value of .05 does not justify concluding that the probability that the samples of participants who got or did not get money are from the same population also equals .05. Therefore, the researcher cannot validly conclude, on this basis, that the null hypothesis is extremely unlikely to be true. But, according to the correlation argument, the fact that p values are correlated with actual probabilities of null hypotheses given findings justifies using the low p value to reject the null hypothesis anyway. If p [or $p(F|H_0)$] and $p(H_0|F)$ are correlated, then low p values generally mean low probabilities of null hypotheses given findings, and so although one does not know the precise probability of the null hypothesis given the finding at hand, the correlation argument suggests that it will be some low number, consequently justifying rejecting the null hypothesis.

Clearly, this argument is only compelling to the degree that $p(F|H_0)$ is correlated with $p(H_0|F)$. For example, suppose the correlation (r) is .9. From this, we can determine that the variance accounted for in $p(H_0|F)$ by $p(F|H_0)$ is 81%. In that case, the utility of $p(F|H_0)$ would be bolstered substantially. Having a procedure that accounts for 81% of $p(H_0|F)$, given that it is rarely possible to obtain an actual value for $p(H_0|F)$, would seem to be an impressive coup for social scientists. In contrast, if the correlation is much less than that, so that $p(F|H_0)$ accounts for only a small percentage of the variance in $p(H_0|F)$, it would be difficult to argue that the ability of $p(F|H_0)$ to predict $p(H_0|F)$ is a good reason for calculating $p(F|H_0)$ values. We recognize that there are other arguments about why $p(F|H_0)$ should or should not be computed; however, our purpose here is to deal with a single issue: the prediction of $p(H_0|F)$.

Method

Overview

To calculate $p(H_0|F)$ through Bayes's theorem, it was necessary to know the following three values: p, which is $p(F|H_0)$; the prior probability of the null hypothesis, $p(H_0)$; and the probability of the finding given that the null hypothesis is not true, $p(F|{-H_0})$ (e.g., see Trafimow, 2003, Equation 2).[1] Although it is easy to obtain $p(F|H_0)$ from experiments (again, this is the same thing as p), the other two values are generally impossible to obtain unless the researcher is willing to make questionable assumptions (for recent discussions, see Trafimow, 2005, 2006). How can the researcher know the prior probability of the null hypothesis? How can the researcher know the probability of the finding given that the null hypothesis is false when there is an infinite number of ways in which it can be false? Fortunately, however, the present goal allowed us to avoid being discommoded by these problems. Because we were interested in the general issue of the correlation between $p(F|H_0)$ and $p(H_0|F)$, we could let $p(F|H_0)$, $p(H_0)$, and $p(F|{-H_0})$ vary between 0 and 1 and use Bayes's theorem to determine $p(H_0|F)$ for each combination of the former three values. To allow as much free play as possible in these three values and avoid giving preferences to particular combinations of them, we assumed that they were uniformly distributed in our computer simulations.[2]

Procedure

The program Visual Basic for Applications (VBA) for Excel (Microsoft Corporation, 2007) was used to generate the 65,000 random data sets. Each data set included a value for $p(F|H_0)$, $p(H_0)$, and $p(F|{-H_0})$ obtained from uniform distributions. For example, the first data set contained the random values .540 [$p(F|H_0)$], .712 [$p(H_0)$], and .185 [$p(F|{-H_0})$]. These values were used in Bayes' formula to derive $p(H_0|F) = .880$. The derived values for $p(H_0|F)$ were then correlated with the randomly generated values for $p(F|H_0)$. In sum, 65,000 random values for $p(F|H_0)$, $p(H_0)$, and $p(F|{-H_0})$ were generated, followed by calculations of $p(H_0|F)$ based on these values, followed by subjecting the random values of $p(F|H_0)$ and the derived values of $p(H_0|F)$ to a Pearson correlation analysis. Subsequently, in the present article, we refer to $p(F|H_0)$ as p_{actual}.

Following these steps, three additional values were generated based on dichotomous cutoff points of significant value or nonsignificant value for $p < .05$, $p < .01$, and $p < .001$. That is, if the p value was significant according to the established cutoff rule, then the cell was assigned a 1; and if it was not significant, then the cell was assigned a 0. Subsequently, in the present article, we refer to these values as $p_{.05}$, $p_{.01}$, or $p_{.001}$, depending on whether the .05, .01, or .001 rule, respectively, was used to establish dichotomous cutoffs for $p(F|H_0)$. The same was done at the .05 cutoff range for $p(H_0|F)$, which we refer to as $p(H_0|F)_{.05}$.

Last, a correlation coefficient was generated for the following comparisons: p_{actual} and $p(H_0|F)$, $p_{.05}$ and $p(H_0|F)$, $p_{.01}$ and $p(H_0|F)$, $p_{.001}$ and $p(H_0|F)$, and $p_{.05}$ and $p(H_0|F)_{.05}$.

Results

The issue of primary interest concerns the ability of p_{actual} to predict $p(H_0|F)$. This correlation (r) was .396. Although the correlation was statistically significant because of the high number of data sets, the variance in $p(H_0|F)$ that p_{actual} accounted for was less than 16%, thereby leaving more than 84% of the variance unaccounted for. In Figure 1, we present a scatterplot of 200 random values from the data set (Figure 1 would simply be a black square if all of the data were represented). This figure clearly illustrates the poorness of the relation between p_{actual} and $p(H_0|F)$.

We also tested some of the dichotomous cutoffs typically used in the social sciences as the criterion for rejecting or retaining null hypotheses. For example, most academic journals set the usual cutoff at $p_{\text{actual}} = .05$, in which case continuous $p(F|H_0)$ values are converted into dichotomous reject–retain decisions. Because this use of

$p(F|H_0)$ values dichotomizes a continuous variable, it would not be surprising if the prediction of $p(H_0|F)$ were affected. In fact, dichotomizing in this way reduces the correlation between $p_{.05}$ and $p(H_0|F)$ to $r = .289$. Now, the variance accounted for decreases to 8.4%, thereby leaving more than 91.6% of the variance unaccounted for.

Journal editors sometimes suggest that lower alpha levels are more scientific (e.g., Melton, 1962). Certainly, lower alpha levels decrease the probability of rejecting the null hypothesis, but at the possible cost of decreasing the prediction of $p(H_0|F)$. Consistent with this possibility, when alpha is set at .01, the correlation between $p_{.01}$ and $p(H_0|F)$ decreases to $r = .188$. Now, the variance accounted for decreases to less than 4%; thus more than 96% of the variance is left unaccounted for. When alpha is set at .001, the correlation between $p_{.001}$ and $p(H_0|F)$ decreases to $r = .089$. Now, the variance accounted for decreases to less than 1%, thereby leaving over 99% of the variance unaccounted for.

Another way of thinking about the association between $p_{.05}$ and $p(H_0|F)$ is in terms of the proportion of times when the two calculations would lead to the same decision or to different decisions assuming an arbitrary cutoff such as 5% for both types of probabilities. To determine this, we dichotomized $p(H_0|F)$ at .05 in a manner similar to that in which we dichotomized $p_{.05}$. Table 1 presents these data as a 2×2 matrix indicating the proportion of cases in which (a) both probabilities were under the cutoff (.05), (b) both probabilities exceeded the cutoff,

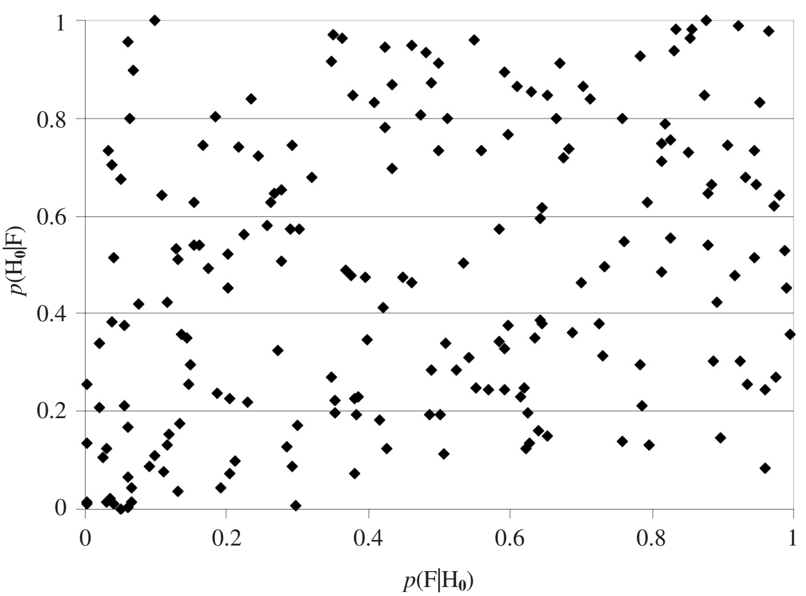

FIGURE 1. Correlation between $p(H_0|F)$ and $p(F|H_0)$.

TABLE 1. Signal Detection Matrix Comparing $p(F|H_0)$ with $p(H_0|F)$, by Proportion

	$p(F\|H_0)$	
$p(H_0\|F)$	Significant	Not significant
Significant	.04	.07
Not significant	.03	.86

(c) $p(H_0|F)_{.05}$ exceeded the cutoff but $p_{.05}$ did not, or (d) $p_{.05}$ exceeded the cutoff but $p(H_0|F)_{.05}$ did not. Clearly, because the cutoff was set at 5%, the computation of both types of probabilities is biased toward retention of the null hypothesis, thereby leading to what might seem to be an impressive level of agreement. However, two alternative methods of correcting for bias suggest that the agreement is not so impressive after all. The simplest method is to compute a correlation (or ϕ coefficient) from the 2×2 matrix. The correlation (r) was .410, and the variance accounted for was 16%, thereby leaving 84% of the variance unaccounted for. The second method is to compute d', which indicates the sensitivity to which decisions based on $p_{.05}$ predict decisions based on $p(H_0|F)_{.05}$. Consistent with the low correlation coefficient, d' is 0.17. In summary, regardless of whether cutoffs are used, $p_{.05}$ does a poor job of predicting $p(H_0|F)_{.05}$.

Discussion

Clearly, $p(F|H_0)$ fails to account for much variance in $p(H_0|F)$, and the variance accounted for decreases substantially when $p(F|H_0)$ is dichotomized by the process of rejecting or retaining the null hypothesis. It is worse that as social scientists reduce the alpha level in an attempt to be more conservative, the variance that is accounted for decreases even further. Therefore, the ability of $p(F|H_0)$ to account for variance in $p(H_0|F)$ does not constitute a compelling justification for their routine use in social science research.

Trafimow (2003) showed that as $p(H_0)$ gets closer to .5, $p(F|H_0)$ exerts a greater influence on $p(H_0|F)$. Thus, the correlation between these two values might increase if the range of values for $p(H_0)$ were restricted to a band around .5; the narrower the band, the greater the correlation. However, even if this were not so, restricting the range of values for $p(H_0)$ would nevertheless increase the correlation because of the following reason. Consider that one of the factors that influences $p(H_0|F)$, other than $p(F|H_0)$, is $p(H_0)$, and so restricting the range of $p(H_0)$ decreases error variance. The decreased error variance naturally increases the size of the correlation between $p(H_0|F)$ and $p(F|H_0)$.

Does this mean that the correlation argument is acceptable after all if $p(H_0)$ is restricted to .5? The following arguments suggest otherwise. Several researchers (most famously, Meehl, 1978; for a review, see Trafimow, 2006) have pointed out that with the current practice of using the point null hypotheses, the null hypothesis is almost always false; given that population parameters can have an almost infinite number of potential values, the probability that any single value is actually precisely correct approaches zero. Consequently, the prior probabilities of null hypotheses, as used by psychologists in actual practice, are much smaller than the optimal value of .5, and so $p(F|H_0)$ would perform even more poorly than it did in our simulations (for a more complete discussion, see Trafimow, 2006).

Of course, it is possible, in principle, to have null hypotheses that specify ranges instead of points. In this case, the prior probabilities of null hypotheses could reasonably have values anywhere between 0 and 1, thereby avoiding the foregoing problem. But this would introduce a second problem: There would be a great deal of variance in $p(H_0)$. From the point of view of the correlation between $p(H_0|F)$ and $p(F|H_0)$, this variance in $p(H_0)$, along with variance in $p(F|–H_0)$, would constitute sources of error variance, and so the implications of the simulations would retain their force.

Clearly, the logic of NHSTP is invalid, and the correlation argument fails to provide a convincing alternative justification. Thus, it is difficult to justify the procedure. This circumstance raises the following question: What should be done? Several suggestions have been put forth. One possibility is to compute the probability of replication (p_{rep}), as advocated by Killeen (2005). The idea of p_{rep} is the following: Suppose a finding in a particular direction is obtained in an experiment. Then, imagine the performance of a second experiment with a similar procedure, the same number of participants, and so on. According to Killeen, p_{rep} indicates the probability that the results of the second experiment would be in the same direction as the results from the first experiment. But, we see two problems with this. First, it is far from clear that the p_{rep} statistic actually is the probability of replication because it does not properly take into account prior probability distributions (e.g., Wagenmakers & Grünwald, 2006). Second, we performed simulations similar to the ones reported in the present article, but using p_{rep} instead of p, and obtained similarly unimpressive correlations with $p(H_0|F)$. This failure is, perhaps, not surprising because p_{rep} is a transformation of p.

Another suggestion is that psychologists use Bayes's theorem, which gives the desired probability of the null hypothesis given the finding. But there are at least three problems with this (for reviews, see Trafimow, 2005, 2006). First, there is no widely agreed upon way to determine the prior probability of the null hypothesis. Second, it is also unclear how to determine the probability of the finding given that the null hypothesis is not true; the null hypothesis may not be true in a variety of ways, thereby complicating the issue. Note that both of these probabilities are necessary to make the Bayesian machinery run. Last—and possibly most important—Bayesian reasoning implies a different definition of probability than the traditional relative frequency or propensity conceptions (Popper, 1983). Bayesians generally are forced to define *probability* in cognitive terms: as a belief state, a value determined by a rational cognitive agent, a value determined by a rational process, or others. The resulting philosophical complications have not been sorted through in a way that induces consensus either among statisticians, scientists, or philosophers, though several attempts have been made.

Trafimow (2005, 2006) proposed an epistemic ratio procedure that has a Bayesian flavor to it but potentially avoids the usual Bayesian problems. A full discussion of his proposal is beyond the scope of the present article, but it retains the traditional relative frequency definition of probability (it is also compatible with a propensity definition; see Popper, 1983). In essence, the idea is for researchers to propose competing hypotheses. The probabilities of the data, given either of the competing hypotheses, are easy to determine (and these are real relative frequency probabilities) and can be compared with each other through an epistemic ratio procedure to aid the researcher in deciding which hypothesis to favor. One disadvantage of the procedure is that it necessitates the formulation of at least two competing hypotheses. Another disadvantage is that it does not actually give the probability of a hypothesis. Rather it provides a way of comparing the relative empirical support for two competing hypotheses. Still, researchers can argue that the disadvantages are relatively minor, at least compared with the disadvantages of alternative procedures.

In conclusion, many previous discussions and reviews have dealt with NHSTP in an attempt to refute all of the arguments that NHSTP proponents have put forth in support of it. We believe that these discussions and reviews are often convincing but have not addressed what we consider to have been a major point in favor of NHSTP: the claim that $p(F|H_0)$ correlates well with $p(H_0|F)$. To our knowledge, the present demonstrations provide the only refutation of this important argument and consequently provide an important piece of the larger picture.

NOTES

1. Equation 2 from Trafimow (2003) is as follows:

$$p(H_0|F) = \frac{p(F|H_0)p(H_0)}{p(F|H_0)p(H_0) + p(F|-H_0)(1 - p(H_0))}$$

   where $p(H_0|F)$ is the probability of the null hypothesis given the finding, $p(F|H_0)$ is the probability of the finding given the null hypothesis, $p(H_0)$ is the prior probability of the null hypothesis, and $p(F|-H_0)$ is the probability of the finding given that the null hypothesis is not true.
2. We also performed simulations with normal distributions to test whether this would improve the prediction of $p(H_0|F)$ from $p(F|H_0)$. In fact, the size of the correlation decreased, possibly because using normal distributions caused a restriction of range.

AUTHOR NOTES

David Trafimow is a professor of psychology at New Mexico State University and an executive editor of *The Journal of General Psychology*. His current research interests include attribution,

attitudes, cross-cultural research, methodology, and potential performance theory. **Stephen Rice** is an assistant professor of psychology at New Mexico State University and a consulting editor of *The Journal of General Psychology*. His current research interests include automation, trust, performance, and aviation psychology.

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. 117–144). 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.

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

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16,* 345–353.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56,* 16–26.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Melton, A. (1962). Editorial. *Journal of Experimental Psychology, 64,* 553–557.

Microsoft Corporation. (2007). *Visual basic for applications* [Computer software]. Redmond, WA: Author.

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. 65–116). Mahwah, NJ: Erlbaum.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Popper, K. R. (1983). *Realism and the aim of science*. London: Routledge.

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

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.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review, 110,* 526–535.

Trafimow, D. (2005). The ubiquitous Laplacian assumption: Reply to Lee and Wagenmakers. *Psychological Review, 112,* 669–674.

Trafimow, D. (2006). Using epistemic ratios to evaluate hypotheses: An imprecision penalty for imprecise hypotheses. *Genetic, Social, and General Psychology Monographs, 132,* 431–462.

Wagenmakers, E.-J., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing. *Psychological Science, 17,* 641–642.

Manuscript submitted March 31, 2008
Revision accepted for publication August 11, 2008