Theory & Psychology
http://tap.sagepub.com/

Why *P* Values Are Not a Useful Measure of Evidence in Statistical Significance Testing

Raymond Hubbard and R. Murray Lindsay

Theory Psychology 2008 18: 69
DOI: 10.1177/0959354307086923

The online version of this article can be found at:
http://tap.sagepub.com/content/18/1/69

Published by:
SAGE
http://www.sagepublications.com

Theory & Psychology

Why *P* Values Are Not a Useful Measure of Evidence in Statistical Significance Testing

Raymond Hubbard
DRAKE UNIVERSITY

R. Murray Lindsay
UNIVERSITY OF LETHBRIDGE

ABSTRACT. Reporting *p* values from statistical significance tests is common in psychology's empirical literature. Sir Ronald Fisher saw the *p* value as playing a useful role in knowledge development by acting as an 'objective' measure of inductive evidence against the null hypothesis. We review several reasons why the *p* value is an unobjective and inadequate measure of evidence when statistically testing hypotheses. A common theme throughout many of these reasons is that *p values exaggerate the evidence against* H_0. This, in turn, calls into question the validity of much published work based on comparatively small, including .05, *p* values. Indeed, if researchers were fully informed about the limitations of the *p* value as a measure of evidence, this inferential index could not possibly enjoy its ongoing ubiquity. Replication with extension research focusing on sample statistics, effect sizes, and their confidence intervals is a better vehicle for reliable knowledge development than using *p* values. Fisher would also have agreed with the need for replication research.

KEY WORDS: likelihood ratios, null hypothesis, (overlapping) confidence intervals, *p* values, posterior probabilities, replication

The most important task before us in developing statistical science is to demolish the P*-value culture, which has taken root to a frightening extent in many areas of both pure and applied science, and technology*. (Nelder, 1999, p. 261)

My personal view is that p*-values should be relegated to the scrap heap and not considered by those who wish to think and act coherently*. (Lindley, 1999, p. 75)

THEORY & PSYCHOLOGY Copyright © 2008 Sage Publications. VOL. 18(1): 69–88
DOI: 10.1177/0959354307086923 http://tap.sagepub.com

Much empirical work in psychology focuses on hypothesis testing. The typical empirical paper develops, tests, and reports the results of a number of explicit hypotheses relating to the topic at hand. The outcomes of these hypothesis tests are said to contribute toward the creation of a body of knowledge within the discipline.

For the most part, psychology researchers rely on *p* values from statistical significance tests when evaluating the merits of their hypotheses. Based on an annual random sample of issues from 12 American Psychological Association journals for the period 1990–2002, for example, Hubbard (2004) estimated that 94% of empirical papers used significance tests.

Given their universality, it seems reasonable to presume that *p* values play an integral part in knowledge development. In addition, the ubiquity of *p* values strongly suggests that researchers are intimately familiar with their capabilities. But this is not always the case. Thus, for instance, many investigators erroneously believe that the *p* value indicates the probability that (1) the results occurred because of chance, (2) the results are replicable, (3) the alternative hypothesis is true, (4) the results are important, and (5) the results will generalize. (For specific examples showing where each of these five misuses of *p* values may be found in the psychology literature, see Bakan, 1966; Carver, 1978; Cohen, 1994; Falk & Greenbaum, 1995; Gigerenzer, 1993; Gigerenzer, Krauss, & Vitouch, 2004; Krämer & Gigerenzer, 2005; Krantz, 1999; Krueger, 2001; Nickerson, 2000; Schmidt, 1996; and Thompson, 1999, among others.)[1]

This paper is not concerned with such misinterpretations of *p* values, damaging though they are. *Rather, it examines the inherent problems associated with the p value as a plausible measure of evidence per se*. Although the origin of the modern *p* value is generally credited to Karl Pearson (1900), who introduced it in his χ^2 test (he actually called it the *P*, χ^2 test), it was Sir Ronald Fisher who was responsible for popularizing statistical significance testing and *p* values in the many editions of his classic books *Statistical Methods for Research Workers* and *The Design of Experiments*, first published in 1925 and 1935, respectively. Fisher used discrepancies in the data to reject the null hypothesis, that is, he calculated the probability of the data on a true null hypothesis, or $\Pr(x \mid H_0)$. Formally, $p = \Pr(T(X) \geq T(x) \mid H_0)$. *P* is the probability of getting a test statistic T(X) greater than or equal to the observed result, T(x), in addition to more extreme ones, conditional on a true null hypothesis, H_0, of no effect or relationship. (Disturbingly, Freund and Perles [1993] remark that differences in the definition of the *p* value abound in textbooks. See Good [1981], also.) So, the *p* value is a measure of the (im)plausibility of the actual observations (as well as more extreme and unobserved ones) obtained in an investigation, assuming a true null hypothesis. The rationale is that if the data are seen as being rare or highly discrepant under H_0, this constitutes *inductive evidence*

against H_0. The idea that rare occurrences comprise evidence against a hypothesis has a pedigree dating back to the first 'significance test' by John Arbuthnot in 1710 concerning the birth rates of males and females in London, and is continued in the work of Mitchell, LaPlace, and Edgeworth, among others. (See Baird, 1988 and Gigerenzer et al., 1989 for synopses of this early history of statistical testing.) Traditionally, a p value of .05 has been used as a benchmark to indicate inductive evidence against the null hypothesis, with values like $p < .01$, $p < .001$, etc., furnishing even stronger evidence against H_0.

Fisher (1959) considered the p value to be an *objective* way for researchers to assess the (im)plausibility of the null hypothesis:

> ... the feeling induced by a test of significance has an objective basis in that the probability statement on which it is based is a fact communicable to and verifiable by other rational minds. The level of significance in such cases fulfils the conditions of a measure of the rational grounds for the disbelief [in the null hypothesis] it engenders. (p. 43)

But a critical question remains: does the p value, in fact, provide an objective, useful, and unambiguous measure of evidence in hypothesis testing? We argue in this paper that it does not. More specifically, a review of the statistics literature points to several reasons—statistical, logical, the relative nature of evidence, etc.—why the p value fails visibly as a credible measure of evidence. Our premise is simple: that p values continue to saturate empirical work is taken as *prima facie* testimony that most psychology (and other) scholars—W. Edwards, Lindman, and Savage (1963), Gigerenzer and his colleagues (Gigerenzer, 1993; Gigerenzer et al., 2004; Gigerenzer & Murray, 1987; Gigerenzer et al., 1989), and Nickerson (2000) being notable exceptions—remain unaware of many of the reasons why this index is a defective measure of evidence. To illustrate this, even those who see value in statistical significance testing (e.g., Abelson, 1997; Chow, 1996, 1998; Cortina & Dunlap, 1997; Hagen, 1997; and Mulaik, Raju, & Harshman, 1997) simply never bring up, much less defend, the issue of the adequacy of the p value as a measure of evidence *qua* evidence. We hope that the present review will help to rectify this situation.

As a secondary goal we propose, like Cohen (1994) and Loftus (1996), that instead of/along with reporting p values in individual studies, researchers should provide estimates of sample statistics, effect sizes, and their confidence intervals. We also stress, following Fisher, the importance of replication with extension research (the grist for meta-analyses) in developing a cumulative knowledge base. For comparisons of population estimates from this research, we recommend the criterion of overlapping confidence intervals. Sufficiently overlapping confidence intervals indicate reasonable estimates of the same population parameter.

Why *P* Values Are an Inadequate Measure of Evidence in Statistical Significance Testing

P Values Exaggerate the Evidence Against the Null Hypothesis

This is the most damning criticism of the p value as a measure of evidence.

Two-sided nulls.
P values exaggerate the evidence against a two-sided (point null or 'small interval') hypothesis (Berger, 1986; Berger & Sellke, 1987). An exact, or point null, hypothesis takes the form H_0: $\theta = \theta_0$ versus H_A: $\theta \neq \theta_0$, where θ_0 is a specific value of θ. More realistically, Berger and Delampady (1987) argue, exact hypotheses are better represented as tests such as H_0: $|\theta - \theta_0| \leq \varepsilon$ versus H_A: $|\theta - \theta_0| > \varepsilon$, where ε is 'small'.

Using a Bayesian significance test for a normal mean, James Berger and Thomas Sellke (1987, pp. 112–113) showed that for p values of .05, .01, and .001, respectively, the posterior probabilities of the null, $\Pr(H_0 \mid x)$, for $n = 50$ are .52, .22, and .034. For $n = 100$ the corresponding figures are .60, .27, and .045. Clearly these discrepancies between p and $\Pr(H_0 \mid x)$ are pronounced, and cast serious doubt on the use of p values as reasonable measures of evidence. In fact, Berger and Sellke (1987) demonstrated that data yielding a p value of .05 in testing a normal mean nevertheless resulted in a posterior probability of the null hypothesis of *at least* .30 for *any* objective (symmetric priors with equal prior weight given to H_0 and H_A) prior distribution.

It is important at this juncture to emphasize the distinction between the p value, $\Pr(x \mid H_0)$, and the posterior probability of the null, $\Pr(H_0 \mid x)$.[2] The p value gives the probability of the observed (and more extreme) data conditional on a true null hypothesis. Even though it may sound similar, this is not the same thing as the probability of the null being true conditional on the observed data. There is an asymmetric relationship between $\Pr(x \mid H_0)$ and $\Pr(H_0 \mid x)$. Despite this, a number of psychologists, including Carver (1978), Cohen (1994), and Nickerson (2000), note that many researchers are confused over the meaning of the two expressions, and tend to view the p value as the probability that the null is true. Berger and Sellke (1987) put this succinctly: 'Indeed, most nonspecialists interpret p precisely as $\Pr(H_0 \mid x)$' (p. 114).

Berger and Sellke's (1987) research led them to conclude that p values can be highly misleading measures of evidence. That is, the use of p values makes it relatively easy to obtain statistically significant findings, such that $p = .05$ can indicate no evidence against H_0. Researchers and practitioners, on the other hand, tend to interpret a .05 value as constituting much greater evidence against the null.

Continuing in the same vein, Berger and Delampady (1987) found similar discrepant results between p values versus $\Pr(H_0 \mid x)$ in both normal and binomial situations. This prompted them to recommend that formal use of p values should be abandoned when testing precise (point null and small interval)

hypotheses, a conclusion supported by Nester (1996). And, of course, psychologists overwhelmingly test point null and small interval hypotheses.

George Casella and Roger Berger (1987), however, showed that Berger and Sellke's (1987) results for two-sided hypotheses do not necessarily extend to the one-sided testing problem. This outcome maintained hope for the efficacy of the *p* value as a measure of evidence, at least in more restricted circumstances. Casella and Berger believe that the *p* value is useful as a quick and crude inferential index. Berger and Sellke (1987) responded:

> Our basic view of the Casella–Berger article, however, is that it pounds another nail into the coffin of *P* values. To clarify why, consider what it is that makes a statistical concept valuable; of primary importance is that the concept must convey a well-understood and sensible message for the vast majority of problems to which it is applied. (p. 135)

Berger and Sellke find no such 'well-understood and sensible message' with respect to *p* values because they do not provide easily interpretable measures of evidence against H_0 over the spectrum of everyday testing problems. Dickey (1987) agreed with Berger and Sellke's position regarding the drawbacks of *p* values, while Dollinger, Kulinskaya, and Staudte (1996) found them wanting even as a measure of evidence for normal data in a one-sided testing context. And in any case, surely science requires more than the quick, crude, restrictive form of inference that Casella and Berger (1987) appear willing to settle for. In light of the above discussion, one would have to concur with Berger and Berry's (1988) sobering opinion that there should be concern about the validity of research findings based on moderately small, including .05, *p* values.

Frequentist 'calibration' of p values.

It is conceivable that the work cited above raising serious doubts on the usefulness of *p* values may be ignored or dismissed by mainstream (Neyman–Pearson) frequentist statisticians because of its 'subjective' Bayesian orientation (see Neyman, 1977). But what if *p* values are found wanting as a measure of evidence among those espousing 'objective' relative frequency approaches to statistical testing. Here, Sellke, Bayarri, and Berger's (2001) findings should serve as a salutary warning even to entrenched (Neyman–Pearson) frequentists. To fully appreciate the importance of this issue requires some background information, supplied below.

It is not understood by many researchers that in classical statistical testing there are two, quite different, measures of 'statistical significance.' One is Fisher's *p* value, which is an inferential index of the strength of the evidence against H_0, is a data-based random variable, and is applicable to individual studies. On the other hand there is the α level from a Neyman–Pearson hypothesis test. This test is concerned with minimizing Type II, or ß, errors (i.e., false acceptance of a null hypothesis) subject to a bound on Type I, or

α, errors (i.e., false rejections of a null hypothesis). In addition, α is a prescription for behaviors (accepting or rejecting H_0), not a means of assessing evidence; is a pre-selected fixed measure, not a random variable; and applies only to long-run repeated random sampling from the same population, not to single experiments (Hubbard, 2004).

The Neyman–Pearson theory of hypothesis testing, with α as the significance level, is generally accepted as constituting frequentist statistical orthodoxy (Hogben, 1957; Nester, 1996; Royall, 1997).[3] So the Neyman–Pearson model is the one typically portrayed in statistics textbooks. Conversely, social science methods texts, in a misguided attempt to present a single, unified model of statistical testing, have tended to anonymously mix together the two incompatible measures of statistical significance, *p*'s and α's. Needless to say, this has resulted in massive confusion among members of the scholarly community about exactly what 'statistical significance' means—is it denoted by a *p* value, an α level, and/or the ubiquitous $p < \alpha$ criterion (Hubbard & Armstrong, 2006)? The upshot is that many researchers (e.g., Bayarri & Berger, 1999, 2000, 2004; Berger, 2003; Berger & Sellke, 1987; Gigerenzer, 1993; Goodman, 1993, 1999; Hubbard, 2004; Hubbard & Bayarri, 2003a, 2003b, 2005) state that *the p value is routinely misinterpreted as a frequentist Type I error probability.*

An empirical literature in which *p* values and α levels are erroneously seen to be interchangeable, but in which investigators overwhelmingly report *p*'s rather than α's required of Neyman–Pearson frequentist orthodoxy (see Hubbard, 2004), sets the backdrop for Sellke et al.'s (2001) study. As seen above, Berger and his colleagues had already shown the *p* value to be a poor gauge of evidence in a Bayesian context. They now wanted to determine if *p* values are useful measures of evidence against H_0 when considered from a Neyman–Pearsonian perspective. Accordingly, Sellke et al. (2001) devised a method for 'calibrating' *p* values so that they can be interpreted as Neyman–Pearson frequentist error probabilities.[4]

The end result of this calibration is as follows: $\alpha(p) = (1 + [-e\,p\log(p)]^{-1})^{-1}$. Consequently, $p = .05$ translates into frequentist error probability $\alpha(.05) = .289$ in rejecting H_0—a result suggesting no evidence against H_0. Even $\alpha(.01) = .111$. These findings convey in a non-Bayesian manner the severe problems involved in using *p* values as credible measures of evidence against the null hypothesis.

Frequentist performance of p values.

As reported in a number of studies (e.g., Berger, 2003; Hubbard & Bayarri, 2003a; and especially Sellke et al., 2001), a simulation of the frequentist performance of *p* values is revealing. Whereas α's can be constrained to some pre-assigned (e.g., .05) level, *p* values share no similar obligation. That is, *p*'s do not behave in the frequentist manner of α's. This is dramatically illustrated by accessing an applet at *www.stat.duke.edu/~berger*, which permits a simulation of the frequentist properties of *p* values.

As an example, suppose we wish to conduct some tests on the effectiveness of a new psychotherapy, P-T. The statistical test would be H_0: P-T = 0 versus H_A: P-T ≠ 0. The simulation, based on a long series of such tests on normal data (variance known), records how often H_0 is true for *p* values in given ranges, say *p* approximately equal to .05 or .01. Otherwise expressed, this frequentist simulation of the behavior of *p* values demonstrates that even when we obtain 'statistically significant' outcomes near the .05 or .01 levels, these results often arise from true null hypotheses of no effect or association. More specifically, assuming that one-half of the null hypotheses in the P-T tests are true, Sellke et al. (2001, p. 63) warned that:

1. Of the subset of P-T tests for which the *p* value is close to the .05 level, *at least* 22% (and typically about 50%) come from true nulls.
2. Of the subset of P-T tests for which the *p* value is close to the .01 level, *at least* 7% (and typically about 15%) come from true nulls.[5]

As Berger (2003) understated the case: 'The harm from the common misinterpretation of $p = 0.05$ as an error probability is apparent' (p. 4). A *p* value of .05 may provide no evidence against the null hypothesis.

P Values and Sample Size

P values and small versus large samples.

Sample size is hugely influential in determining significance levels. Royall (1986), for example, cites well-known statisticians whose interpretations of *p* values in small versus large sample studies are totally contradictory: some argue that a given *p* value in a small sample study is stronger evidence against H_0 than the same *p* value in a large sample study, and vice versa. As such, a given *p* value does not have a fixed, objective meaning—it is contingent upon (at least) the sample size. Indeed, as Marden (2000) points out, the *p* value is not very useful with large sample sizes. Because almost no null hypothesis is *exactly* true (Tukey, 1991), when sample sizes are large enough almost any null hypothesis will have a tiny *p* value. Hand's (1998) concerns about the relevance of significance testing in data-mining situations, where every *p* value will be statistically significant to several zeros following the decimal point, is simply Marden's observation writ bold.

Lindley's 'paradox'.

Moreover, the problems with *p* values and sample sizes do not end here. We must consider also Lindley's 'paradox' (Lindley, 1957). He showed that for any level of significance, *p*, and for any nonzero prior probability of the null hypothesis, $Pr(H_0)$, a sample size can be found such that the posterior probability of the null, $Pr(H_0 \mid x)$, is $1 - p$. That is, a null hypothesis that is soundly *rejected* at, say, the .05 level by a Fisherian significance test can nevertheless have 95%

support from a Bayesian viewpoint. That these inferences are diametrically opposed is the paradox. The rationale behind this conundrum, Johnstone (1986) explains, is that no matter how small the *p* value, the likelihood ratio $Pr(x \mid H_0)/Pr(x \mid H_A)$ approaches infinity as the sample size increases. Consequently, for large *n*, a small *p* value can actually be interpreted as evidence *in favor of* H_0 rather than *against* it. The question of the objectivity and usefulness of the *p* value as a measure of evidence is shattered by this argument.

Experimental Designs and P Values

How different investigators might conceive the planning and execution of a study can also lead to *p* values with widely varying magnitudes. As an example of this, let us examine Fisher's (1935, ch. 2) classic experiment of the 'lady tasting tea,' as described by Lindley (1993). The lady in question claimed she could distinguish between whether milk or tea had been poured *first* into a cup (of tea). In the experiment, the lady is presented with six pairs of cups of tea, and she must determine whether milk or tea entered the cup first. The null hypothesis—that she cannot, in fact, discriminate—is that she would guess 50% right (R) and 50% wrong (W). Suppose that she gets the first five results right and the last one wrong, or RRRRRW. The *p* value for this outcome, Lindley notes, is $7(½)^6$, or .110, which is not statistically significant at the .05 level. This *p* value, like all of them, consists of two parts. In this case: $6(½)^6 = .094$ (probability of observed outcomes) $+ (½)^6 = .016$ (probability of more extreme outcomes). The justification for the inclusion of the latter in the calculation of *p* values is given in a later section of the paper.

Suppose instead of the above design, another researcher decides to repeat the experiment until the lady makes her first mistake. In this case, and with the same RRRRRW data, the *p* value is now statistically significant at the .032 level [$(½)^6 + (½)^6 = .016 + .016 = .032$]. The two parts of this *p* value are explained as follows: $(½)^6 = .016$ (probability of observed outcomes)—but without this expression being multiplied by 6 because the mistaken choice, W, must always come at the end (see, e.g., Goodman, 1999)— $+ (½)^6 = .016$ (probability of more extreme outcomes).

Of course, these experimental results make no sense. The exact same data, obtained in the exact same sequence, should yield the exact same *p* values. But they do not. And all because two different investigators held alternate conceptions as to how the experiment should be run.

Effect Sizes and P Values

One must surely question the *p* value as a measure of evidence when it has nothing to say about the *effect size* obtained in a study (Gelman & Stern,

TABLE 1

Trial	No. preferring A	No. preferring B	% preferring A
1	15	5	75.0
2	114	86	57.0
3	1,046	954	52.3
4	1,001,455	998,555	50.07

2006). For instance, a small sample study with a large effect can yield the same *p* value as a large sample study with a small effect size. To illustrate this, consider Freeman's (1993) hypothetical medical trials in which all patients receive both treatments A and B and are asked to express their preference (see Table 1).

The results of trial 1, with its 75% preference rate for A over B, would be considered as indicative of a potentially enormous preference for A. Trial 4, on the other hand, with a 50.07% preference rate for A, would be regarded as overwhelming evidence that preferences for A versus B are all but identical. Very few researchers would view the results of these four trials as being equivalent, yet they all produce a *p* value of .041. (Freeman does not specify which particular statistical test was used in making these comparisons.)

Gibbons' (1986) assertion, therefore, in an article titled '*P*-Values', that 'An investigator who can report only a *P* value conveys the maximum amount of information contained in the sample…' (p. 367) is seen to be incredulous. Far from conveying such information, Berger, Boukai, and Wang (1997) caution that the interpretation of *p* values will change drastically from problem to problem. Contrary to Fisher's claims, the *p* value is not an objective measure of evidence against a hypothesis, a topic that is pursued below.

P Values and Subjectivity

A further example of the fallibility of the *p* value as an objective measure of evidence is seen in the choice of one-sided versus two-sided statistical significance tests (Goodman & Royall, 1988; Royall, 1997). Although two-sided tests are the norm, researchers are sometimes advised that if they expect a departure from H_0 in a specific direction they can halve the *p* value, say from .05 to .025. That is, Goodman and Royall (1988) comment, even though the data are the same, the *p* value is altered by the researcher's subjective impressions about the likely outcome of the study. They also note that similar changes to *p* values occur when the research involves multiple comparisons.

P Values Are Logically Flawed

P values are logically flawed measures of support for hypotheses.
Schervish (1996) demonstrated that *p* values fail to meet the simple logical condition required by a measure of support, namely, that if hypothesis H implies hypothesis H′, there should be at least as much support for H′ as there is for H. In the course of this work, he lamented that he had been unable to construct a consistent interpretation of the *p* value as anything resembling a measure of support for a hypothesis even in simple, much less multiparameter, problems. Schervish warned that 'common informal use of *P* values as measures of support or evidence for hypotheses has serious logical flaws' (p. 203). Further, because they are not as different as they might have seemed (i.e., point null and one-sided hypotheses are, in fact, at opposite ends of a continuum of hypotheses spanned by interval hypotheses), Schervish argued that the interpretation of the *p* value as a measure of evidence should be consistent across the different hypotheses tested—point null, one-sided, and interval. This, of course, is not the case. Thus, Schervish's research supports the claim of Berger and Sellke (1987) and Bayarri and Berger (2000) that the *p* value is not amenable to a reasonably objective interpretation as evidence over the spectrum of testing problems. And this, together with much other information presented in this paper, runs counter to Frick's (1996) claim that a *p* value creates a common measure of strength of evidence across statistical tests.

The p value computes not the probability of the observed data under H_0, but this plus the probability of more extreme data.
This is a major weakness regarding the usefulness of *p* values. Because they are defined as a procedure for establishing the probability of an outcome, *as well as more extreme ones,* on a null hypothesis, significance tests are affected by how the probability distribution is spread over unobserved outcomes in the sample space. That is, the *p* value denotes not only the probability of what was observed, but also the probabilities of all the more extreme events that did not arise. How is it that these more extreme, unobserved, cases are involved in calculating the *p* value? To find out, we revisit Lindley's (1993) analysis of the 'lady tasting tea'.

Recall the lady was right (R) about the outcomes of the first five experiments, and wrong (W) about the sixth, i.e., RRRRRW. This result has probability $(½)^6$, or a statistically significant *p* value of .016. But, Lindley continues, Fisher saw the flaw in this argument because *every* possible result with the six pairs of cups is significant at $p = .016$. To guard against this, Fisher proposed that *any* result where just one W occurs out of six supports the lady's ability to discriminate, and should be included in the calculation of the *p* value. There are six possibilities, including RRRRRW, so the *p* value is now $6(½)^6 = .094$, which is not significant.

Fisher's significance rationale is no longer the *p* value for a given outcome on a true null hypothesis, but that and similar outcomes; in our case, one mistake in six taste tests. He was aware, however, that this situation also was not feasible. This is because the most likely result in the which comes first—milk or tea—taste test is sheer guessing: 50% R and 50% W. For example, Lindley asserts, for 128 taste tests (64 R, 64 W) the *p* value—$^{128}C_{64}(½)^{128}$—is approximately .05. But this brings us back to square one; if this result is the most likely, then all other outcomes have a smaller probability. That is, all 128 taste tests will be significant at the $p = .05$ level.

In order to circumvent this issue, Fisher suggested that if one error in six (RRRRRW) is significant, more *extreme* outcomes, such as no mistakes at all (RRRRRR), must necessarily be significant. Therefore, these more extreme results should be incorporated when calculating the *p* value. For the outcome RRRRRW, with probability $(½)^6$ or $p = .016$, there are five others (RRRRWR, RRRWRR, etc.) as extreme, and one (RRRRRR) more extreme, so the overall probability is $7(½)^6 = .110$, which is not significant. And this *p* value has two components: $6(½)^6 = .094$ (probability of observed data) *plus* $(½)^6 = .016$ (probability of more extreme data). This $p = .110$ is, of course, the same value cited earlier.

Many statisticians (e.g., Berger & Berry, 1988; Berger & Delampady, 1987; Freeman, 1993; Goodman, 1999; Royall, 1997) charge that a valid measure of strength of evidence cannot be dependent on the probabilities of unobserved outcomes. Jeffreys (1939) acknowledged this illogic in *p* values:

> *What the use of* P *implies … is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred*. This seems a remarkable procedure. (p. 316)

Royall (1997) insists that there is no value to Fisherian significance tests because they are at odds with the law of likelihood and its implication of the 'irrelevance of the sample space' (p. 68). As he explains:

> The law of likelihood says that the evidence in an observation, $X = x$, as it pertains to two probability distributions labeled θ_1 and θ_2, is represented by the likelihood ratio, $f(x;\ \theta_1)/f(x;\ \theta_2)$. In particular, the law implies that for interpreting the observation as evidence for hypothesis H_1: $\theta = \theta_1$ *vis-à-vis* H_2: $\theta = \theta_2$, only the likelihood ratio is relevant. What other values of X might have been observed, and how the two distributions in question spread their remaining probability over the unobserved values is irrelevant—all that counts is the ratio of the probabilities of the observation under the two hypotheses. (p. 22)

Or as Freeman (1993) says, echoing Birnbaum's (1962) and A.W.F. Edwards' (1992) seminal contributions:

> … the likelihood principle is the one secure foundation for all statistics. I find the arguments in favour of it compelling and the counterarguments

> unconvincing. Since *p*-values and all other frequentist methods violate this principle, they must necessarily be unsatisfactory tools of statistical inference. (pp. 1444–1445)[6]

Specification of an Alternative Hypothesis

Evidence is relative.

When an alternative hypothesis is specified, it is possible to identify those outcomes as extreme or more so than the observed event. Consequently, Royall (1997) states, it is not low probability under A that makes an observation evidence against A. Rather, it is low probability under A compared with the probability under a different hypothesis B, and this makes it evidence against A versus B. This line of reasoning necessitates a weighing of the evidence between two rival hypotheses, a situation impossible in Fisherian significance tests, where there is only the null hypothesis. Fisher never saw the need for an alternative hypothesis, and vigorously opposed its later inclusion by Jerzy Neyman and Egon Pearson (Gigerenzer & Murray, 1987; Hubbard & Bayarri, 2003a).

Note, then, Johnstone's (1986) observation that the law of likelihood provides a better measure of evidence than *p* values for evaluating the plausibility of two (or more) rival hypotheses.[7] More specifically, if the likelihood ratio $Pr(x \mid H_0)/Pr(x \mid H_A)$ exceeds 1, then the evidence is in favor of H_0 over H_A, and vice versa. *Unfortunately, Fisher's disjunction only addresses* $Pr(x \mid H_0)$*; it is silent about* $Pr(x \mid H_A)$. The *p* value is a tail-area probability and not a likelihood ratio.

Our interest is in the alternative hypothesis.

Specifying an alternative hypothesis is not just a means of covering values more extreme than those observed on a null hypothesis. The alternative (research) hypothesis is the one the investigator is interested in. Berkson (1942) recognized this when posing an early challenge to Fisher's paradigm of null hypothesis testing:

> In the null hypothesis schema we are trying only to nullify something …. But ordinarily evidence does not take this form. With the corpus delicti in front of you, you do not say, 'Here is evidence against the hypothesis that no one is dead'. You say, 'Evidently, someone has been murdered'. (p. 326)

For statistical tests to be scientifically useful they should speak to the research hypothesis, and not be fixated with rejection of the null hypothesis. This is consistent with Goodman and Royall's (1988) complaint that *p* values blinker us into thinking that a hypothesis can only be weakened, rather than strengthened, by the data. But Fisher's methodology denies the existence of an alternative/research hypothesis. In this matter, it is sometimes thought that Fisherian significance testing has an implicit alternative hypothesis that is

simply the complement of the null. But, as Hubbard and Bayarri (2003a) point out, this is difficult to formalize. For instance, what is the complement of an N (0,1) model? Is it the mean differing from 0, the variance differing from 1, the model not being Normal? Formally, Fisher only had the null model in mind, and wanted to see if the data were compatible with it.

Confidence Intervals, Not *P* Values

The foregoing discussion makes it clear that *p* values are neither objective nor credible measures of evidence in statistical significance testing. Moreover, the authenticity of many published studies with $p < .05$ findings must be called into question. Rather than the preoccupation with *p* values and testing, the goal of empirical research in individual studies should be the estimation of sample statistics, effect sizes, and the confidence intervals (CIs) surrounding them. CIs underscore the superiority of estimation over testing. Scientific advance typically necessitates plausible estimates of the magnitude of effect sizes in the population (A.W.F. Edwards, 1992; Lindsay, 1995), and the CI provides this. CIs also indicate the precision or reliability of the estimate via the width of the interval. Also, because they are in the same metric as the point estimate, CIs make it easier to see whether the results are substantively, rather than statistically, significant. And, of course, a CI can be used as a significance test; a 95% CI not including the null value is equivalent to rejecting the hypothesis at the .05 level.

Furthermore, initial results need to be replicated and extended. Here again, CIs assume a pivotal role. Specifically, we advocate the criterion of *overlapping CIs* around point estimates across similar studies as a measure of replication success. Substantially overlapping CIs suggest tenable estimates of the same population parameter, and we applaud the very useful recent work in this area (e.g., Cumming & Finch, 2001, 2005; Fidler, Thomason, Cumming, Finch, & Leeman, 2005; Goldstein & Healy, 1995; Schenker & Gentleman, 2001; Schmidt, 1996; Smithson, 2003; Thompson, 2002; Tryon, 2001).[8]

It is the *systematic* replication and extension of the results of previous studies, and not *p* values from individual ones, that fosters *cumulative* knowledge development. That this statement appears to have eluded many applied researchers, as well as editors and reviewers, is puzzling because Fisher (1966) himself put only provisional stock in statistically significant results from single studies: 'we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon' (p. 13).

Fisher was a major proponent of replication: 'Fisher had reason to emphasize, as a first principle of experimentation, the function of appropriate *replication* in providing an estimate of error' (Fisher Box, 1978, p. 142). Indeed, Fisher Box (1978) insinuates that Fisher coined the term 'replication': 'The method adopted

was replication, as Fisher called it; by his naming of what was already a common experimental practice, he called attention to its functional importance in experimentation' (p. 142). Fisher (1966) encouraged in particular the importance of replication *with extension* research: 'we may, by deliberately varying in each case some of the conditions of the experiment, achieve a wider inductive basis for our conclusions, without in any degree impairing their precision' (p. 102). It is easy, therefore, to imagine Fisher agreeing with the sentiments put forward in both the psychology (e.g., Falk, 1998; Hubbard, 2004; Neuliep & Crandall, 1990, 1993; Rosenthal, 1990; Rosnow & Rosenthal, 1989; Sohn, 1998; Thompson, 1994) and statistics (e.g., Bayarri & Mayoral, 2002; Chatfield 1995; Ehrenberg & Bound, 1993; Guttman, 1977; Lindsay & Ehrenberg, 1993; Nelder, 1986, 1999; Ottenbacher, 1996; Rosenbaum, 1999, 2001) disciplines that there is an urgent need for more replication with extension research.

Conclusions

Over the last few decades a considerable literature has emerged in psychology critical of the misuse of statistical significance testing. Much of the literature has dealt with how researchers invest these tests with far greater capabilities than they possess. Moreover, this frequently involves gross misinterpretations of the meaning of p values. Works like these are to be welcomed. During this same time, however, little has appeared in psychology (or elsewhere in the social sciences) about the severe limitations of the p value as a measure of evidence *per se.* In other words, it is bad enough for researchers to misuse a measure that is useful: But it strains credulity to do so when that measure is seriously flawed in itself. And this paper has demonstrated—from a multitude of perspectives—that the p value is just that. Hence Nelder's (1999) call to 'demolish' the p value culture.

In concluding, we note that there is more than a hint of irony in the fact that Fisher's sanctioning of the vital role of replication has been overlooked, while at the same time his widely misunderstood and defective p values blanket the empirical literature. This has occurred, even though, as Steiger (1990) expressed: 'An ounce of replication is worth a ton of inferential statistics' (p. 176). It is past time to redress this imbalance. Accordingly, we hope that the present paper will help stimulate further public discussion on methods of data analysis and knowledge development within the field.

Notes

1. These works, particularly Nickerson's (2000) *tour de force*, also offer excellent reviews of the statistical significance testing controversy in psychology.
2. From Bayes' theorem, the posterior probability of the null hypothesis using our terminology is calculated as follows:

$$\Pr(H_0|x) = \frac{\Pr(x|H_0)\Pr(H_0)}{\Pr(x|H_0)\Pr(H_0) + \Pr(x|H_A)\Pr(H_A)}$$

Readers are referred to several articles in the psychology literature making use of this formula (e.g., Cohen, 1994; Falk & Greenbaum, 1995; Hagen, 1997; Nickerson, 2000).

3. Fisher is also a frequentist in the sense that a *p* value of .05 on a true null hypothesis yielded in a single study would be interpreted to mean that the probability of obtaining such an observed value (and more extreme ones) is only 5%. He is not, however, a frequentist in the long-run repeated sampling mode like Neyman–Pearson. See Hubbard and Bayarri (2003a) for further discussion of this.
4. The details of this calibration are too involved to consider here. They can be found in Sellke et al. (2001).
5. Interested readers are encouraged to experiment with the applet, where one can specify the initial percentage of true nulls, the small ranges of *p* values to investigate (e.g., $p = .05$ might be chosen as *p* between .049 and .05), and the value of the normal means, μ's, that occur under H_A in the simulation.
6. Freeman's (1993) appraisal of the usefulness of *p* values in data analysis is instructive, reflecting as it does a 180º change of opinion:

> This paper started life as an attempt to defend *p*-values … I have, however, been led inexorably to the opposite conclusion, that the current use of *p* values as the 'main means' of assessing and reporting the results of clinical trials is indefensible. (p. 1443)

7. See also Glover and Dixon's (2004) support of the likelihood principle as a means of adjudicating knowledge claims in psychology.
8. Despite the advantages in using CIs over *p* values, reforms in statistical practice in psychology have been problematic (Hubbard & Ryan, 2000). Fidler, Thomason, Cumming, Finch, and Leeman (2004), for example, report on the difficulties encountered by Loftus (1993) in his efforts to decrease the emphasis on significance testing while editor of *Memory & Cognition.* During his tenure, Fidler et al. (2004) note, the proportion of articles using error bars (both CIs and standard error bars) increased to 41% as compared with 7% under his predecessor. Unfortunately, after Loftus left his editorial position, this proportion fell to 24%. Clearly, effecting changes in the manner in which statistical evidence is presented in the literature will be no easy task. Yet it is surely an important one.

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.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117–141). Mahwah, NJ: Erlbaum.

Baird, D. (1988). Significance tests, history and logic. In S. Kotz & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences* (pp. 466–471). New York: Wiley.

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

Bayarri, M.J., & Berger, J.O. (1999). Quantifying surprise in the data and model verification (with comments). In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian statistics* (Vol. 6, pp. 53–82). Oxford: Clarendon.

Bayarri, M.J., & Berger, J.O. (2000). *P* values for composite null models. *Journal of the American Statistical Association*, *95*, 1127–1142.

Bayarri, M.J., & Berger, J.O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, *19*, 58–80.

Bayarri, M.J., & Mayoral, A.M. (2002). Bayesian design of 'successful' replications. *The American Statistician*, *56*, 207–214.

Berger, J.O. (1986). Are *p*-values reasonable measures of accuracy? In I.S. Francis, B.F.J. Manly, & F.C. Lam (Eds.), *Pacific Statistical Congress* (pp. 21–27). Amsterdam: Elsevier.

Berger, J.O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? (with comments). *Statistical Science*, *18*, 1–32.

Berger, J.O., & Berry, D.A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*, 159–165.

Berger, J.O., Boukai, B., & Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis (with comments). *Statistical Science*, *12*, 133–160.

Berger, J.O., & Delampady, M. (1987). Testing precise hypotheses (with comments). *Statistical Science*, *2*, 317–352.

Berger, J.O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *p* values and evidence (with comments). *Journal of the American Statistical Association*, *82*, 112–139.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, *37*, 325–335.

Birnbaum, A. (1962). On the foundations of statistical inference (with comments). *Journal of the American Statistical Association*, *57*, 269–326.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378–399.

Casella, G., & Berger, R.L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with comments). *Journal of the American Statistical Association*, *82*, 106–139.

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference (with comments). *Journal of the Royal Statistical Society A*, *158*, 419–466.

Chow, S.L. (1996). *Statistical significance: Rationale, validity and utility.* Thousand Oaks, CA: SAGE.

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

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

Cortina, J.M., & Dunlap, W.P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161–172.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532–574.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*, 170–180.

Dickey, J.M. (1987). Comment on Berger and Sellke. *Journal of the American Statistical Association*, *82*, 129–130.

Dollinger, M.B., Kulinskaya, E., & Staudte, R.G. (1996). When is a *p*-value a good measure of evidence? In H. Rieder (Ed.), *Robust statistics, data analysis and computer intensive methods* (pp. 119–134). New York: Springer Verlag.

Edwards, A.W.F. (1992). *Likelihood* (Expanded ed.). Baltimore, MD: Johns Hopkins University Press.

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

Ehrenberg, A.S.C., & Bound, J.A. (1993). Predictability and prediction (with comments). *Journal of the Royal Statistical Society A*, *156*, 167–206.

Falk, R. (1998). Replication—A step in the right direction: Commentary on Sohn. *Theory & Psychology*, *8*, 313–321.

Falk, R., & Greenbaum, C.W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*, 75–98.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*, 119–126.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2005). Still much to learn about confidence intervals: Reply to Rouder and Morey (2005). *Psychological Science*, *16*, 494–495.

Fisher, R.A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver & Boyd.

Fisher, R.A. (1935). *The design of experiments.* Edinburgh: Oliver & Boyd.

Fisher, R.A. (1959). *Statistical methods and scientific inference* (2nd ed.). Edinburgh: Oliver & Boyd.

Fisher, R.A. (1966). *The design of experiments* (8th ed.). Edinburgh: Oliver & Boyd.

Fisher Box, J. (1978). *R.A. Fisher: The life of a scientist.* New York: Wiley.

Freeman, P.R. (1993). The role of *p*-values in analysing trial results. *Statistics in Medicine*, *12*, 1443–1452.

Freund, J.E., & Perles, B.M. (1993). Observations on the definition of P-values. *Teaching Statistics*, *15*, 8–9.

Frick, R.W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379–390.

Gelman, A., & Stern, H. (2006). The difference between 'significant' and 'not significant' is not itself statistically significant. *The American Statistician*, *60*, 328–331.

Gibbons, J.D. (1986). *P*-Values. In S. Kotz & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences* (pp. 366–368). New York: Wiley.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C.A. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: SAGE.

Gigerenzer, G., & Murray, D.J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The empire of chance.* New York: Cambridge University Press.

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, *11*, 791–806.

Goldstein, H., & Healy, M.J.R. (1995). The graphical interpretation of a collection of means. *Journal of the Royal Statistical Society A*, *158*, 175–177.

Good, I.J. (1981). Some logic and history of hypothesis testing. In J.C. Pitt (Ed.), *Philosophy in economics* (pp. 149–174). Dordrecht: D. Reidel.

Goodman, S.N. (1993). *P* values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, *137*, 485–496.

Goodman, S.N. (1999). Toward evidence-based medical statistics. 1: The *P* value fallacy. *Annals of Internal Medicine*, *130*, 995–1004.

Goodman, S.N., & Royall, R.M. (1988). Evidence and scientific research. *American Journal of Public Health*, *78*, 1568–1574.

Guttman, L. (1977). What is not what in statistics. *The Statistician*, *26*, 81–107.

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

Hand, D.J. (1998). Data mining: Statistics and more? *The American Statistician*, *52*, 112–118.

Hogben, L. (1957). *Statistical theory.* New York: Norton.

Hubbard, R. (2004). Alphabet soup: Blurring the distinctions between *p*'s and α's in psychological research. *Theory & Psychology*, *14*, 295–327.

Hubbard, R., & Armstrong, J.S. (2006). Why we don't really know what *statistical significance* means: Implications for educators. *Journal of Marketing Education*, *28*, 114–120.

Hubbard, R., & Bayarri, M.J. (2003a). Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing (with comments). *The American Statistician*, *57*, 171–182.

Hubbard, R., & Bayarri, M.J. (2003b). *P* values are not error probabilities. *Institute of Statistics and Decision Sciences*, *Working Paper*, *No. 03–26.* Durham, NC: Duke University Working Papers Series, 27708–0251.

Hubbard, R., & Bayarri, M.J. (2005). Comment on Christensen. *The American Statistician*, *59*, 353.

Hubbard, R., & Ryan, P.A. (2000). The historical growth of statistical significance testing in psychology—and its future prospects. *Educational and Psychological Measurement*, *60*, 661–681.

Jeffreys, H. (1939). *Theory of probability.* Oxford: Clarendon.

Johnstone, D.J. (1986). Tests of significance in theory and practice (with comments). *The Statistician*, *35*, 491–504.

Krämer, W., & Gigerenzer, G. (2005). How to confuse with statistics or: The use and misuse of conditional probabilities. *Statistical Science*, *20*, 223–230.

Krantz, D.H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *44*, 1372–1381.

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

Lindley, D.V. (1957). A statistical paradox. *Biometrika*, *44*, 187–192.

Lindley, D.V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*, 22–25.

Lindley, D.V. (1999). Comment on Bayarri and Berger. In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian Statistics* (Vol. 6, p. 75). Oxford: Clarendon.

Lindsay, R.M. (1995). Reconsidering the status of tests of significance: An alternative criterion of adequacy. *Accounting, Organizations and Society*, *20*, 35–53.

Lindsay, R.M., & Ehrenberg, A.S.C. (1993). The design of replicated studies. *The American Statistician*, *47*, 217–228.

Loftus, G.R. (1993). Editorial comment. *Memory & Cognition*, *21*, 1–3.

Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171.

Marden, J.I. (2000). Hypothesis testing: From *p* values to Bayes factors. *Journal of the American Statistical Association*, *95*, 1316–1320.

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

Nelder, J.A. (1986). Statistics, science and technology (with comments). *Journal of the Royal Statistical Society A*, *149*, 109–121.

Nelder, J.A. (1999). From statistics to statistical science (with comments). *The Statistician*, *48*, 257–269.

Nester, M.R. (1996). An applied statistician's creed. *The Statistician*, *45*, 401–410.

Neuliep, J.W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality*, *5*, 85–90.

Neuliep, J.W., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior and Personality*, *8*, 22–29.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, *36*, 97–131.

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

Ottenbacher, K.J. (1996). The power of replications and replications of power. *The American Statistician*, *50*, 271–275.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, *50*, 157–175.

Rosenbaum, P.R. (1999). Choice as an alternative to control in observational studies (with comments). *Statistical Science*, *14*, 259–304.

Rosenbaum, P.R. (2001). Replicating effects and biases. *The American Statistician*, *55*, 223–227.

Rosenthal, R. (1990). Replication in behavioral research. *Journal of Social Behavior and Personality*, *5*, 1–30.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284.

Royall, R.M. (1986). The effect of sample size on the meaning of significance tests. *The American Statistician*, *40*, 313–315.

Royall, R.M. (1997). *Statistical evidence: A likelihood paradigm.* New York: Chapman & Hall.

Schenker, N., & Gentleman, J.F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, *55*, 182–186.

Schervish, M.J. (1996). *P* values: What they are and what they are not. *The American Statistician*, *50*, 203–206.

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

Sellke, T., Bayarri, M.J., & Berger, J.O. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician*, *55*, 62–71.

Smithson, M. (2003). *Confidence intervals.* Thousand Oaks, CA: SAGE.

Sohn, D. (1998). Statistical significance and replicability: Why the former does not presage the latter. *Theory & Psychology*, *8*, 291–311.

Steiger, J.H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180.

Thompson, B. (1994). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality*, *62*, 157–176.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology*, *9*, 165–181.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 25–32.

Tryon, W.W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371–386.

Tukey, J.W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*, 100–116.

ACKNOWLEDGEMENTS**.** We have benefited from discussions and correspondence on this topic with Susie Bayarri, Jim Berger, and Rahul Parsa. We would also like to thank two anonymous reviewers and the editor for suggestions that have improved the manuscript. Any errors and shortcomings are our responsibility.

RAYMOND HUBBARD is Thomas F. Sheehan Distinguished Professor of Marketing in the College of Business and Public Administration, Drake University. His research interests include applied methodology and the sociology of knowledge development in the management and social sciences. He has published a number of articles on these topics. ADDRESS: College of Business and Public Administration, Drake University, Des Moines, IA 50311, USA. [email: Raymond.Hubbard@drake.edu]

R. MURRAY LINDSAY is Dean of the Faculty of Management and Professor of Accounting, University of Lethbridge, Alberta. He has a special interest in developing a theory of replication and generalization, and examining the role modern statistical methods play within such a theory. He has published several articles in these areas. ADDRESS: Dean, Faculty of Management, University of Lethbridge, Lethbridge, AB T1K 3M4, Canada. [email: m.lindsay@uleth.ca]