Ann. Zool. Fennici 46: 311–349 ISSN 0003-455X (print), ISSN 1797-2450 (online)
Helsinki 30 October 2009

Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian

Stuart H. Hurlbert[1] & Celia M. Lombardi[2]

[1)] *Department of Biology, San Diego State University, San Diego, California 92182, U.S.A. (e-mail: shurlbert@sunstroke.sdsu.edu)*
[2)] *Consejo Nacional de Investigaciones Científicas y Técnicas, Museo Argentino de Ciencias Naturales, Av. Angel Gallardo 470, C1405DJR Buenos Aires, Argentina (e-mail: celia7@sigmaxi.net)*

Received 26 Jan. 2009, revised version received 5 June 2009, accepted 13 July 2009

Hurlbert, S. H. & Lombardi, C. M. 2009: Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. — *Ann. Zool. Fennici* 46: 311–349.

This essay grew out of an examination of one-tailed significance testing. One-tailed tests were little advocated by the founders of modern statistics but are widely used and recommended nowadays in the biological, behavioral and social sciences. The high frequency of their use in ecology and animal behavior and their logical indefensibility have been documented in a companion review paper. In the present one, we trace the roots of this problem and counter some attacks on significance testing in general. Roots include: the early but irrational dichotomization of the *P* scale and adoption of the 'significant/non-significant' terminology; the mistaken notion that a high *P* value is evidence favoring the null hypothesis over the alternative hypothesis; and confusion over the distinction between statistical and research hypotheses. Resultant widespread misuse and misinterpretation of significance tests have also led to other problems, such as unjustifiable demands that reporting of *P* values be disallowed or greatly reduced and that reporting of confidence intervals and standardized effect sizes be required in their place. Our analysis of these matters thus leads us to a recommendation that for standard types of significance assessment the paleoFisherian and Neyman-Pearsonian paradigms be replaced by a neoFisherian one. The essence of the latter is that a critical α (probability of type I error) is not specified, the terms 'significant' and 'non-significant' are abandoned, that high *P* values lead only to suspended judgments, and that the so-called "three-valued logic" of Cox, Kaiser, Tukey, Tryon and Harris is adopted explicitly. Confidence intervals and bands, power analyses, and severity curves remain useful adjuncts in particular situations. Analyses conducted under this paradigm we term neoFisherian significance assessments (NFSA). Their role is assessment of the existence, sign and magnitude of statistical effects. The common label of null hypothesis significance tests (NHST) is retained for paleoFisherian and Neyman-Pearsonian approaches and their hybrids. The original Neyman-Pearson framework has no utility outside quality control type applications. Some advocates of Bayesian, likelihood and information-theoretic approaches to model selection have argued that *P* values and NFSAs are of little or no value, but those arguments do not withstand critical review. Champions of Bayesian methods in particular continue to overstate their value and relevance.

"*… the object of statistical methods is the reduction of data. A quantity of data … is to be replaced by relatively few quantities which shall adequately represent the whole ….*" (Fisher 1922)

"*The decade of the 1990's has been a critical one in hypothesis testing's protracted struggle for survival. During this decade especially vitriolic attacks, by especially viable attackers, in especially visible outlets … have been mounted for the greater good of God, country, and no significance testing*." (Levin 1998a)

"*… use [of null hypothesis significance testing] has been explicitly denounced by most eminent and most experienced scientists, both on theoretical and methodological grounds ….*" (Lecoutre *et al*. 2001)

"*Some academicians and philosophers have proposed that the social science community ban the use of statistical significance tests. Their declarations are not worth taking seriously. Getting beyond the rants about the limitations of conventional significance tests is important ….*" (Boruch 2007)

Introduction

Misuse of one-tailed tests

Early versions of this essay had as a subtitle, "Lessons learned from an analysis of one-tailed testing", as it grew out of an analysis of misuse and misprescription of one-tailed statistical tests in the natural, behavioral, and social sciences (Lombardi & Hurlbert 2009). That analysis examined the historical literature on the subject, criticized the vague, illogical and inconsistent advice usually offered by statistics books on it, and assessed frequency of use of one-tailed tests in the 1989 and 2005 volumes of two journals, *Animal Behaviour* and *Oecologia*. Averaged over the two years, one-tailed tests were used in > 24% of *Animal Behaviour* articles and > 13% of *Oecologia* articles. Synthesizing the cogent arguments of a few earlier writers on the subject (Kimmel 1957, Goldfried 1959, Pillemer 1991, Harris 2005), we concluded that one-tailed tests are rarely appropriate in basic or applied research in any discipline. A few textbooks have adopted this viewpoint (e.g. Welkowitz *et al*. 1971, 1991, 2006, Fleiss 1981, 1986, Altman 1991, Schulman 1992, Bart *et al*. 1998), but the great majority have not.

Thus, of the hundreds of one-tailed tests found in our survey, none was or could be justified logically. How did such inappropriate procedures become so widespread? Proximately, it surely has been the fault of the poor advice on the topic found in most statistics texts — and, presumably, in university courses that use those texts. But where did those texts and professors get their ideas?

The present article had its genesis in an attempt to answer that question. Close reading of the historical literature located three primary origins for the confusion. First are deficiencies in the decision theoretic framework that has dominated statistical practice in the English-speaking world from, roughly, the 1940s. That framework in its current form is "a mixture of [the methods of] R. A. Fisher, on the one hand, and Jerzy Neyman and Egon S. Pearson on the other, a mixture that none of these statisticians (certainly not Neyman and Pearson) would have approved" (Gigerenzer & Murray 1987, Gigerenzer *et al*. 2004). After a dissection of its history and incongruities, Salsburg (1993; *see* also Salsburg 1992) concluded: "And so the Neyman-Pearson formulation lays in rubble at our feet. It is an arbitrary construction with no apparent relationship to the needs of clinical research." Or, we would add, to any other type of research. That framework has promoted rigid and simplistic thinking by dichotomization of the scale of *P* values. It has also promoted the notion that a high *P* value constitutes evidence favoring the null over the alternative hypothesis.

A second source of confusion has been claims that researchers overemphasize *P* values and present effect sizes and confidence intervals too infrequently. If true this would be simply bad science and bad writing and not a weakness of statistical

methods themselves. Combined with the illogic of a dichotomized *P* scale, these ideas have stimulated some critics to propose that reporting of *P* values should be forbidden and only confidence intervals and effect sizes be allowed.

Third, the distinction between statistical hypotheses and scientific or research hypotheses has become blurred in the minds and writings of many scientists and statisticians. That confusion has fostered use of one-tailed tests when researchers have not understood that a significance test or assessment is just a procedure for summarizing certain properties of a particular data set and not by itself a test of a *scientific* hypothesis.

The proposed framework shift

The decision theoretic framework as presented, without that label, in Neyman and Pearson (1933a, 1933b) is summarized in Fig. 1, some version of which now appears in most introductory statistics textbooks. The italicized phrase *illogical decision* is our own modification of it.

This framework built on Fisher's (1925) ideas of the null hypothesis and fixed critical *P* values for assessing statistical significance. To those, Neyman and Pearson added the concepts of α as a long term (type I) error rate, of accepting the null hypothesis when $P > \alpha$, and of alternative hypotheses, power and type II error (Neyman & Pearson 1933a, 1933b). The framework was further codified and expanded by Wald's (1939, 1950) work on decision theory and then by Lehmann's (1959, Lehmann & Romano 2005) classic treatise. But technically the two frameworks are fundamentally incompatible. Lenhard (2006) gives an excellent analysis of the "profound conceptual basis" for the distractingly polemical battles between Fisher and Neyman and Pearson, which was that "both sides held conflicting views about the function of mathematical models and about the role of modelling in statistical inference."

We will advocate discarding the Neyman-Pearson framework for most significance testing situations and replacing it with an explicitly neoFisherian one that (1) does not fix α; (2) does not describe *P* values as 'significant' or 'non-significant'; (3) does not accept null hypotheses on the basis of high *P* values but only suspends judgment; (4) interprets significance tests in accord with the "three-valued logic" of Cox (1958), Harris (1997a) and others; (5) recognizes the obvious, near universal need to present effect size information in conjunction with significance tests; and (6) acknowledges the frequent utility of confidence intervals (and other adjunct statistics helpful to interpretation) as well as the fact that they are often unneeded. Procedures carried out under this paradigm we refer to as neoFisherian significance assessments (NFSA). This label may help diminish the notion that they constitute tests of scientific hypotheses. It also distinguishes them from null hypothesis significance tests or testing (NHST), a label that can be retained for procedures carried out under paleoFisherian or Neyman-Pearsonian paradigms. As we discuss later, Neyman and Pearson seem not to have intended their framework to be used *in scientific work* in the rigid manner adopted in many quarters, and, by the end of their careers might have found what we term the neoFisherian paradigm quite reasonable.

	Reality	
Decission	H_0 true	H_0 false
Accept H_0 when $P > \alpha$	correct conclusion *illogical decision*	Type II error probability = β *illogical decision*
Reject H_0 when $P \leq \alpha$	Type I error probability = α	correct conclusion probability = $1 - \beta$ = power

Fig. 1. The Neyman-Pearson decision theoretic framework. We have inserted "illogical decision" in two of the boxes to emphasize that high *P* values do not constitute evidence in favor of H_0. Hence acceptance of H_0, or preference for it over H_1, is never a logical decision, and type II errors are risked only by those who do not understand this fact.

For anyone contemplating adoption of the neoFisherian framework, an immediate concern is how it would affect their own writing. 'How does this system actually work?', some reviewers of this paper asked. There is a foreboding sense the shift would be seismic, draconian, but it would not be. Adoption leads to simpler, clearer, more natural writing. Perhaps that is why neoFisherian manuscripts have been sneaking into the literature 'under the radar' of paleoFisherian and Neyman-Pearsonian editors for years.

Here, for example, is a 'random' selection of ecological papers all of which use significance tests of one type or another and all of which conform to the neoFisherian paradigm in their interpretations and language: Greenwald and Hurlbert (1993), Hart *et al*. (1998), Timms (1998), Detwiler *et al*. (2002), Caskey *et al*. (2007), Henny *et al*. (2007), Moreau *et al*. (2007), Reifel *et al*. (2007), Sardella *et al*. (2007), Swan *et al*. (2007), Sockman (2008), Lombardi and Hurlbert (2009). These provide examples of nuanced interpretation of *P* values and show just how dispensable 'α' and 'significant' are.

Dichotomization of the *P* scale

$P \leq \alpha$; tails? who cares!

The popularity of one-tailed tests (Burke 1953, Siegel 1956, Fleiss 1987, Peace 1989, 1991, Zar 2004, Lombardi & Hurlbert 2009) is due, in part, to the idea that significance testing requires specification of α, a fixed critical *P* value (Simon 1986, Freedman *et al*. 1991). This dichotomization of the *P* scale and our language interferes with the clear presentation and judicious interpretation of statistical analyses. Large numbers of statisticians and scientists have strongly criticised this dichotomization, a fact of which most researchers and authors of statistics textbooks seem unaware. The notion that critical *P* values must be specified has more to do with personalities, unclear writing, and accidents of history than with logic or utility.

Freedman *et al*. (1991) noted "It is the arbitrary lines at 5% and 1% which make the distinction between two-tailed and one-tailed tests loom so large. There is no sharp dividing line between probable and improbable results. A *P* value of 5.1% means just about the same thing as 4.9%. However, these two *P* values can be treated quite differently, because many journals will only publish results that are 'statistically significant' — the 5% line. Some of the more prestigious journals will only publish results which are 'highly significant' — the 1% line."

A prime example: an editor of the *Journal of Experimental Psychology* once explained, "In editing the *Journal* there has been a strong reluctance to accept and publish results related to the principal concern of the research when those results were significant [only] at the 0.05 level, *whether by one- or two-tailed test* [our emphasis]" (Melton 1962). Such policies and attitudes on the part of editors and manuscript reviewers have provided strong incentive to the use of one-tailed tests by naive researchers. Why risk manuscript rejection with your *P* value of 0.08 from a two-tailed test, when you can quietly make a *post hoc* 'prediction', redo your test as a one-tailed one, and obtain $P = 0.04$?

Old roots

The idea of fixing α and dichotomizing the scale of *P* values has old roots (Hogben 1957, Cowles & Davis 1982, Hall & Selinger 1986, Gigerenzer & Murray 1987, Cowles 1989, Salsburg 1992, Huberty 1993, Inman 1994). By the early 19th century some scientists were quantifying the 'spread' of the normal distribution by the probable error. This is equal to 0.6746 times the standard deviation though the latter term was coined much later (Pearson 1894). In a normal distribution, 50% of the observations lie within one probable error of the mean.

Dichotomization of terminology followed development of the concept of probable error. Venn (1866) referred to the concept of significance, and in a later work (Venn 1888: 147) stated, "When we are dealing with statistics, we ought to be able not merely to say vaguely that the difference does or does not seem significant to us, but we ought to have some test as to what difference would be significant." W. S. Gosset (Student 1908) suggested that a deviation from the mean "three times the probable error in the normal curve, for most purposes, would be considered significant." In this he was almost certainly following the lead of Karl Pearson whose lectures at University College London Gosset had attended in 1906. In his unpublished lecture notes, Karl refers to deviations greater than three probable errors as "definitely significant." But Karl did not think in terms of hard dichotomies and also spoke of "not definitely," "possibly," "probably," and "almost certain[ly]" significant (Stigler 2000, 2005, Ziliak & McCloskey 2008).

The paleoFisherian paradigm

The tail areas of the curve beyond the mean plus or minus three probable errors sum to 4.56%. In his influential manual, *Statistical Methods for Research Workers*, Fisher (1925) rounded this 4.56% to 5% when publishing the first tables of z, t, and χ^2 values corresponding to specific critical P values. He recommended, in connection with the z test, that it was "convenient to take this point [$P = 0.05$] as a limit in judging whether a deviation is to be considered significant or not" (p. 47). He also stated that "in practice we do not want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion …. We shall not often be astray if we draw a conventional line at .05 and consider that higher values of χ^2 indicate a real discrepancy" (p. 79). By the 13th edition, this last sentence had been changed to: "A [χ^2] value exceeding the 5 per cent point is seldom to be disregarded" (Fisher 1958: 80). This focus on 5% notwithstanding, six of the seven tables in Fisher (1925) gave test statistic values corresponding to several other P values (Stigler 2008).

In his other influential book, *The Design of Experiments*, Fisher (1935a: 15) stated, "It is usual and convenient for experimenters to take 5 per cent as a standard level of significance, in the sense they are prepared to ignore all results which fail to reach this standard …." This statement persisted through all editions of that book.

Null hypothesis testing has no inherent requirement that an α be specified or that the 'significant/non-significant' terminology be adopted. Fisher may have been impelled to those conventions, however, not only by historical antecedents but also by a very practical and personal obstacle. Kendall (1963) relates that: "He [Fisher] himself told me that when he was writing *Statistical Methods for Research Workers* he applied to Pearson for permission to reproduce Elderton's table of chi-squared and that it was refused. This was perhaps not simply a personal matter because the hard struggle which Pearson had for long experienced in obtaining funds for printing and publishing statistical tables had made him most unwilling to grant anyone permission to reproduce. He was afraid of the effect on sales of his *Tables for Statisticians and Biometricians* [K. Pearson 1914] on which he relied to secure money for further table publication. It seems, however, to have been this refusal which first directed Fisher's thoughts towards the alternative form of tabulation with quantiles as argument, a form which he subsequently adopted for all his tables and which has become common practice."

This is what Fisher referred to when he explained the absence from his book of more extended tables as "owing to copyright restrictions" (Fisher 1925: 78, 1958: 79). Fisher did not invent the 'significant/non-significant' dichotomy, but his books and novel tabulations of critical values of test statistics played a large role in its rapid and wide dissemination.

Enshrinement of the dichotomy was completed when Neyman and Egon Pearson (1933a) adapted the ideas of significance testing to create their decision theoretic framework. In contrast to Fisherian significance testing, this approach *requires* the specification of α and is suited for situations, such as industrial quality control or "commercial specifications" (Neyman & Pearson 1933b), where different actions will be taken according to whether $P \leq \alpha$ or $P > \alpha$. Where the weighing of evidence, and not the taking of an action on the basis of a single significance test, is called for, as in virtually all basic and applied research, the specification of α is superfluous. It only incites inappropriately dichotomous language and thinking. But since Fisher, Neyman and E. Pearson disagreed on other statistical matters, often vehemently and publicly, the fact that they *seemed* to agree on the need to specify α has carried great weight. Since that time, recipe-hungry researchers, editors, and textbook writers have taken α specification as an obligatory step in the carrying out of significance tests. Such specification and a refusal to accept H_0 when $P > \alpha$ are core elements of the paleoFisherian paradigm of significance testing.

It is ironic that, though Fisher's (1925) tabulation of critical P values was driven by practical considerations, publication of those tables shortly before Neyman and Pearson began their collaboration in 1926 was a key stimulus to Neyman and Pearson basing their framework on an obligatory *a priori* setting of α (Lehm-

ann 1993). Moreover, the proposition of a clear evidentiary standard such as $\alpha = 0.05$ may have been a critical and "brilliant stroke of simplification that opened the arcane domain of statistical calculation to a world of experimenters and research workers" who were confronting the new statistical methodologies with some trepidation (Stigler 2008).

There were additional ironies. Neyman and Pearson's (1933a) paper putting forward the decision theoretic framework was communicated to the Royal Society of London in 1932 by Karl Pearson "who was hostile and skeptical of its contents" (Reid 1982: 103), possibly because it furthered the notion of a need to specify α. Karl was not only E. Pearson's father, he also developed the χ^2 test and applied it in the first systematic use of significance testing. Karl, however, did *not* specify α or critical P values, in carrying out his tests. A few years later he (Pearson 1935b) said it was "unwise" of Fisher (1925) to have created "tables which provide only the [critical test statistic] values of [for] $P = 0.01$ and $P = 0.05$," as in the expanded z table in Fisher (1930). Karl Pearson further stated that "The value of P at which we consider goodness or badness of graduation [i.e. conformity or disconformity with H_0] starts cannot be fixed without regard to the special problem under consideration." In other words, calculate P, but perhaps there is no need to specify α.

Unfortunately this message did not carry. Karl Pearson, then two years into retirement, was not only justifiably going against Fisher, Neyman, and E. Pearson on α specification, he was also justifiably under fire from Fisher (1935b) and Buchanan-Wollaston (1935) for use of χ^2 as a "goodness-of-fit" test, i.e. as a way of confirming null hypotheses or of measuring the "goodness of graduation" (Inman 1994). Significance tests can never confirm or provide evidence in favor of null hypotheses. Debates where each side is half right and half wrong often lead the audience astray. The "goodness-of-fit" label was and is a highly misleading one. It probably bears some responsibility for decades of misuse and misinterpretation of χ^2 and related tests. "Badness-of-fit" would be a label more consonant with what χ^2, like other signficance tests, actually assesses.

Some other early textbook writers suggested that exact P values be reported (Hagood 1941: 447–451, Mather 1951: 21). None went so far, however, as to state explicitly that specification of α could therefore be dispensed with.

The neoFisherian paradigm

Toward the end of his life Fisher came close to admitting that K. Pearson was correct, that the specification of α was superfluous, and that the reporting of exact P values was desirable, albeit often still difficult in the 1950s, well before the advent of statistical software packages. In his last and more philosophical book, Fisher (1956) said, "no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects [null] hypotheses; he rather give his mind to each particular case in the light of his evidence and ideas." This thought was also incorporated into the last editions of his two more influential books. In the 13th edition of *Statistical Methods for Research Workers*, Fisher (1958: 128) inserted the statement that "tests of significance are used as an aid to judgment, and should not be confused with automatic acceptance tests or 'decision functions'." In the 7th edition of *The Design of Experiments*, Fisher (1960) added a new, one-page section titled "Scientific Inference and Acceptance Procedures." This contrasted the use of significance tests for weighing of evidence with their use in acceptance procedures *a la* Neyman and Pearson. In that section, Fisher stated (p. 25), "Convenient as it is to note that a hypothesis is contradicted at some familiar level of significance such as 5% or 2% or 1%, we do not, in Inductive Inference, ever need to lose sight of the exact strength [i.e. exact P value] which the evidence has in fact reached …." Ziliak and McCloskey (2008: 232) aware of Fisher's 1956 statement but not those of 1958 and 1960, reject the idea that there had been a genuine evolution of Fisher's thought. They claim, without evidence, that Fisher was only "playing a game … fearful of losing influence" to other statisticians.

Those added statements were perhaps the most fundamental changes made over the many editions of these two books. But they were brief,

belated and somewhat cryptic, and, in any case, by 1958 much damage had been done. Large numbers of statistics books were on the market, and the great majority had adopted the hybrid Neyman-Pearson-paleoFisherian decision framework and its requirement of α specification.

The force of these text changes was also diminished by Fisher's retention, in the seventh and subsequent editions of *The Design of Experiments*, of a statement contradicting them. This was the "usual and convenient" statement quoted earlier. That statement implies that if α has been set at 0.05, then it is proper to "ignore" a result associated with a P value of 0.06. That was the *only* prescription that Fisher offered his readership for the 25 years preceding the 7th edition of *The Design of Experiments* and 33 years preceding the 13th edition of *Statistical Methods for Research Workers*.

Clearer distillation of the arguments against α specification appeared about the same time as Fisher's near recantation. Cox (1958: 366–368) offered what might be regarded as the first précis of the neoFisherian paradigm, opposing "rigid dividing line[s]" and also advocating, with Yates (1951), more attention to effect sizes. Eysenck (1960) argued bluntly against the specification of an α and against use of the terms "significant" and "insignificant" (or 'non-significant'). He pointed out that these habits have "no obvious advantage," usually no logical rationale, and tend to lead to "gross absurdities" and increasingly convoluted language ("almost significant," "significant at the 10% level," etc.), as when we set $\alpha = 0.05$ and obtain two P values of 0.04 and 0.06, respectively. He argued for reporting exact P values and taking the interpretation from there. Skipper *et al*. (1967) reviewed the issue and concluded that, "Tradition notwithstanding, there seems to be little justifiable reason [for dichotomizing our interpretation of the P scale and recommended that] scientists … do away with arbitrary levels of significance, and the calling of one test result 'significant' and another 'not significant'."

In his detailed retrospective analysis of R. A. Fisher's works, Savage (1976) observed that "[a]pparently there have been statisticians who recommended actually picking a level [of α] before an experiment and then rejecting or not according as that level was obtained. I do not have the impression that any professional statisticians make that recommendation today, though it is still often heard among those who are supposed to be served by statistics …." We agree with the implication that researchers in the social and natural sciences in the 1970s, as now, were using the inappropriate decision-theoretic framework and terminology. We are less certain, however, about the blamelessness of "professional statisticians" in promoting that framework, then or now. Huberty (1993) reviewed 57 statistics textbooks written for the behavioral sciences between 1910 and 1992 and found that since the 1950s the fixed-α approach has strongly dominated. We assume that many of these authors were or are "professional statisticians." On the other hand, Salsburg (1992: 26) claimed that, "Philosophically, the English school [of statistics] continues to follow Fisher, using significance tests as a relatively vague and rough cutting tool, where there is no predetermined level of significance that signifies action or nonaction." As English statistician Altman (1991: 168–169) puts it, "It is ridiculous to interpret the results of a study differently according to whether the P value obtained was, say, 0.055 or 0.045. These P values should lead to very similar conclusions, not diametrically opposed ones …. In recent years there has been a welcome move away from regarding the P value as significant or not significant, according to which side of the arbitrary 0.05 value it is, towards quoting the actual P value …. Forcing a choice between significant and non-significant obscures the uncertainty present whenever we draw inferences from a sample."

Journal editorial boards and other textbook authors are still stumbling on this issue. Current instructions for contributors to *Animal Behaviour* (2009), for example, specify that α shall be set at 0.05 in absence of specific justification for another value, and that "Nonsignificant outcomes should be indicated with an exact probability value wherever possible, or as NS or $P > 0.05$, as appropriate for the test." Gotelli and Ellison (2004: 97–98) throughout their text emphasize the supposed need to establish "the precise cutoff point [α] that we should use in making the decision to reject or not reject the

null hypothesis" and, in two hypothetical examples (p. 329, 334) imply that reporting only "$P < 0.05$" or "ns" is adequate. But they also suggest that "in many cases, it may be more important to report the exact *P*-value and let the readers decide for themselves how important [perhaps *certain* is intended here] the results are."

From discussion of this issue with other scientists, it seems the biggest psychological impediment to the acceptance of the neoFisherian paradigm is a reluctance to throw out that deceptive crutch, the phrase 'statistically significant'. As Stoehr (1999) points out, we all would like a "quick, objective and automatic way" to evaluate our results, but there is none that also meets the additional requirements of 'logical' and 'useful'. We must simply apply the same sorts of nuanced thinking and nuanced language we use in other contexts involving gradations in strength of evidence.

If the critics of the idea that α must be specified had been widely heeded, which they clearly have not been, natural and social scientists would have been spared decades of misleading, sleep-inducing reportage of statistical analyses in language of the decision theoretic framework. Though many modern statisticians and scientists have recommended the reporting of exact *P* values and less foolishness over the mystical α (e.g., Cox 1958, 1977, 2006a, Gibbons & Pratt 1975, Henkel 1976, Barnard 1982, Altman *et al*. 1983, Mead & Curnow 1983, Yates 1984, Moore 1985:323, Ware *et al*. 1986, Gardner & Altman 1989, Rosnow & Rosenthal 1989, Camilli 1990, Daniel 1990, Freedman *et al*. 1991, Wilkinson & TFSI 1999, Salsburg 1992, 1993, Huberty 1993, Wang 1993, Frick 1996, Rossi 1997, Stoehr 1999, Kline 2004, Christensen 2005, Hubbard & Armstrong 2006, Fidler *et al*. 2006), editors, textbook writers, and researchers have yet to give them much credence.

A core principle of this neoFisherian paradigm, then, is that in testing situations, an α should not be specified, and terms such as 'statistically significant' and 'statistically non-significant' should not be used, nor should useless and misleading symbolic notation such as 'ns' and '$P > 0.05$'. The neoFisherian label seems appropriate for three reasons. First, Fisher clearly was moving toward this position at the end of his career. Second, his original conception of significance testing did not require specification of a critical *P* value even though he appended that superfluity to it for reasons essentially psychological, historical and accidental in nature. And third, other concepts formalized by Neyman and Pearson but that we regard as admissible under the neoFisherian paradigm — such as alternative hypotheses, power, and confidence intervals — were all implicit in Fisherian significance testing regardless of what Fisher said about them or of how unsuccessful his idea of "fiducial intervals" proved to be. Spanos (1999: 561) notes, "As far as testing is concerned …. Fisher's procedure has not been superseded by that of Neyman and Pearson as the traditional treatment [in textbooks] would have us believe." A neoFisherian approach should overcome some of the principal objections of that majority of critics of significance testing who have mistakenly assumed α specification to be an obligatory component of such procedures.

Although Gigerenzer *et al*. (2004) advance the unsupportable proposition that "Only when one knows extremely little about a topic … might a null hypothesis test be appropriate," they at least do not recommend banning them. More favorably, they do advocate the neoFisherian position and strongly recommend against teaching the paleoFisherian or Neyman-Pearsoninan paradigms or hybrids of them. Likewise, though they also say that significance tests have only "marginal value," Hubbard and Armstrong (2006) suggest that when they are employed, this should be done in a neoFisherian manner. They also usefully emphasize the critical distinction between α and *P*, the fact that a *P* value cannot be interpreted as an 'observed' α, and the mostly ignored fact that "α plays no role in Fisherian significance … [and] the *p* value plays no role in N-P [Neyman-Pearson] tests" once the accept/reject decision demanded by that paradigm is made. Pollard and Richardson (1987), Goodman (1999a) and Hubbard and Bayari (2003) give good clarifications of the α-*P* distinction also.

There are at least three books that in different ways advocate the neoFisherian paradigm. One, an advanced text on mathematical statistics, is actually titled *Principles of statistical inference from a neo-Fisherian perspective* (Pace & Salvan 1997). Spanos's (1999) magnum opus

is admirable for its thoroughness, clarity and implicit advocacy of the neoFisherian paradigm, although, ironically, Ziliak and McCloskey (2008: 107) congratulate Spanos for "trying to crack the Fisher monopoly on advanced econometrics." We do, however, chide Spanos (1999) in a later section for its misinterpretation of high *P* values. The book's scope is broader than its subtitle might seem to imply and deserves the attention of all disciplines. The third book, Cox (2006a), also deserves a wide audience. It is an elegant extended essay by one of the world's pre-eminent statisticians and is more focused on the philosophical and logical issues that are the subject of this article, with emphasis on comparison of Bayesian and frequentist methods. All three books advocate for most scientific work the reporting of exact *P* values and the superfluity of fixed α's, and, by implication, the inappropriateness of the 'significant/non-significant' terminology. Perhaps they presage and provide a foundation for a new generation of statistics textbooks.

Mellowing of Neyman and Pearson

Neyman and E. S. Pearson both seem to have recognized that the formal rigidity of their framework was not well suited to scientific research. In her reformulation of the Neyman-Pearson paradigm, Mayo (1992, 1996: 377–395, 407–411), building on the detective work and suggestions of Birnbaum (1977), gives a convincing analysis of how Pearson retreated early. The "hints and suggestions" in E. S. Pearson's published and unpublished writings, from his first paper with Neyman (Neyman & Pearson 1928) to a much later reflective essay (Pearson 1955) indicate that Pearson in fact "rejected the statistical philosophy that ultimately became associated with NP statistics." He seems to have recognized that "evidential" rather than "behavioral" (i.e. accept/reject) interpretations of test results made more sense in scientific work, and that there should be flexibility in interpreting *P* values $\geq \alpha$. Shades of Karl Pearson's 1905–1906 lecture notes! Was Egon honoring his father while also becoming a closet neoFisherian?

There is good evidence in his papers of the 1950s that Neyman also had second thoughts (Mayo & Spanos 2006). We have pointed out (Lombardi & Hurlbert 2009) that although "Salsburg (1992: 23–24) claimed that Neyman never championed the [decision theoretic] framework after the mid-1930s and seemingly 'agreed in principle with most of Fisher's criticisms' of it," Neyman (1950, 1976) did champion the framework in its standard form in his textbook and a later essay. Recognizing the appropriateness of 'evidential' rather than 'behavioral' (Birnbaum 1977) interpretations of significance tests in scientific as opposed to industrial contexts, Neyman may have become a 'situational' statistician, defending one position in his theoretical writings and adopting a more flexible one in empirical science contexts. In 1964, Neyman wrote to E. S. Pearson, "The time when I was a theoretician is past. Now it's either galaxies, or cell division, or carcinogenesis, etc." (Reid 1982: 267) — and, as it proved shortly, cloud seeding for rain production. In a later philosophical essay, Neyman (1977: 112) recounted their cloud-seeding studies, and labeled *P* values of 0.09, 0.03, and < 0.01 reported in their earlier paper (Lovasich *et al.* 1971), as "approximately significant," "significant," and "highly significant," respectively. The dichotomies of the paleoFisherian and Neyman-Pearsonian frameworks were quietly admitted to be less appropriate than more nebulous interpretations — at least in cloud work! Indeed, Cox (2006a: 43, 195) has noted that "the differences between Fisher and Neyman … were not nearly as great as the asperity of the arguments between them might suggest … [and in] actual practice … Neyman … often reported *p*-values whereas some of Fisher's use of tests … was much more dichotomous"!

A brave last gasp

A strange argument for fixing α values is found in Frick's (1996) otherwise excellent defense of null hypothesis testing when appropriately used. Frick recommends that α not only be fixed but that it be fixed at 0.05. He thus defies many modern journals and textbooks which suggest that it can be set higher or lower than 0.05 according to some usually vaguely defined 'specifics of the situation' or postulated relative costs of type I and

type II errors. Referencing earlier authors, Frick counters that "there is little reason for experimenters to choose different levels of alpha … [as] two different experimenters should not reach different statistical conclusions given the same data … it is appropriate that alpha is set by the enterprise of psychology." Or, perhaps, he might suggest, by the "enterprise" of science as a whole?

Frick's clearly presented reasoning is valid only if we accept his premise that some α always needs to be specified and that the 'significant/non-significant' terminology is to be used. In that case, a result yielding $P = 0.07$ *would* likely be described and interpreted differently by two researchers, one who sets $\alpha = 0.05$ and another who sets $\alpha = 0.10$.

Once the superfluous, paleoFisherian-Neyman-Pearsonian premise is disposed of, however, the problem disappears, and the two researchers would likely come to more similar conclusions. Similar, not identical, of course, for individuals will always vary as to what they might term 'tentative', 'moderate', 'strong', or 'very strong' evidence against H_0.

The larger question Frick (1996) poses, and answers, is, "Should the experimenter decide what amount of evidence is sufficient for a finding to enter the corpus of psychology? Obviously not." We agree. Editors and referees — and hopefully a wider readership — can and will make their own varied decisions as to what weight to accord that $P = 0.07$ and will do so more judiciously in the absence of any prescribed α. When all other criteria determining the value of a study to "the corpus of pyschology" are taken into account, good editors and referees will find some results with $P = 0.07$ more valuable and publishable than others with $P = 0.02$. So we agree with Frick that "allowing the experimenter to select alpha is unneeded and inappropriate." Let's just also not give that authority to editors, editorial boards, and all Higher Level Committees! They have no grounds for demanding specification of α, let alone the setting of it at any particular value. Ziliak and McCloskey (2008: 249) concur: "No uniform minimum level of Type I error should be specified or enforced by journals, governments, or professional associations."

Acceptance of null hypotheses

A second defect of the decision theoretic framework is the notion that when $P > \alpha$, this is evidence that H_0 is true and should be 'accepted' or 'retained'. That interpretation of high P values is neither a logical nor a necessary component of significance testing. It is, however, a notion present in the original formulation of the framework (Neyman & Pearson 1933b), and it represents one of the most widespread misinterpretations of significance tests by scientists (Grant 1962, Henkel 1976, Oakes 1986, Sedlmeier & Gigerenzer 1989, Cohen 1990, Altman 1991, Inman 1994, Schmidt 1996, Hurlbert 1998, Johnson 1999, Marden 2000, Tryon 2001, Hurlbert & Lombardi 2003, Balluerka *et al*. 2005, Levine *et al*. 2008a, b). Dar *et al*. (1994), for example, found that of 163 psychotherapy studies published during 1967-1988, 36% interpreted "[N]onsignficant results of … ANOVAs … to mean that the groups were statistically equivalent." Of 200 articles published in *Ecology* and *Journal of Ecology* in 2001–2002 or 2005 that reported a "non-significant result", Fidler *et al*. (2006) noted that 47% in 2001–2002 and 63% in 2005 interpreted $P > \alpha$ as "no effect." Krueger (2001) claimed that "reasoning pragmatically, most researchers" accept H_0 when $P > \alpha$. Taylor and Gerrodette (1993) warned how bad management decisions could result from acceptance of null hypotheses in conservation biology studies. This could occur if these are interpreted in terms of "the dominant paradigm for hypothesis testing … [that] involves a yes/no decision about the falsity of the null hypothesis" and no account is taken of the often low power of such studies deriving from small sample sizes, the infeasibility of manipulative experiments, and other constraints.

An early discussion in Nature

The logical fallacy involved in accepting H_0, or retaining it but not H_1 also, on the basis of a high P value, though codified by Neyman and Pearson (1933a, 1933b), was not original with them. It doubtless evolved in close conjunction with the idea of a dichotomized P scale, K. Pearson's

and W. S. Gosset's (Student 1906–1907, Berkson 1942) notions about 'goodness-of-fit' tests, and Gosset's (Student 1908) original misinterpretation of P as the probability that H_0 is true. As early as 1931 even statistics texts had begun stating, "The [null] hypothesis is accepted if the level is fairly high and … if the level is low (say below 0.05) the hypothesis is rejected" (Tippett 1931: 69–70).

An early discussion of the fallacy was provided in a series of letters by Buchanan-Wollaston (1935), Fisher (1935b), and K. Pearson (1935a, 1935b) in *Nature* (Inman 1994). Buchanan-Wollaston raised the issue giving it as one reason for distrust by Continental statisticians and scientists of British statistical procedures. Fisher replied that "Mr. Buchanan-Wollaston's point that [tests of significance are] cogent for the rejection of [null] hypotheses but not for their acceptance, deserves to be widely appreciated," and made the further valid point that "'errors of the second kind' [Type II errors] are committed only by those who misunderstand the nature and application of tests of significance." A person can voluntarily risk making a Type II error, i.e. risk accepting H_0 or retaining only it when it is false, but such a risk is never compelled or argued for by the outcome of a significance test. If P is high and one decides that H_0 cannot be rejected, one suspends judgment, and the only error possibly committed is that of having wasted resources on a study that provides no firm basis for choosing between H_0 and H_1 or for stating the direction (sign) of an effect. Such a study, of course, may provide useful information relative to statistical hypotheses other than the original one tested. For example the hypothesis that μ_A is at least 50% greater than μ_B could be confidently rejected on the basis of the data in Table 1.

Unfortunately this simple message on the illogic of allowing a P value ever to drive the acceptance of H_0 did not come through this scrambled Fisher-Pearson exchange. The letters addressed several other issues and were used by Fisher and Pearson to throw barbed comments at each other. More importantly, in the final letter of the exchange, Pearson (1935b) misread Fisher's unclear prose and accused him of a "logical fallacy" whereas the two men almost certainly were in agreement on the point at issue. Fisher (1935b) had stated that "tests of significance, when used accurately … are never capable of establishing [null] hypotheses as certainly true." Pearson (1935b) countered that significance tests do not establish the truth or falsity of hypotheses, but only can provide evidence bearing on them. Pearson also correctly stated that "if an hypothesis is false, its reverse must be true," i.e. that if we reject H_0, we accept H_1, at least when H_1 is the standard composite alternative hypothesis of $\partial \neq 0$. Where formal assessment of such additional hypotheses might be useful, there is an abundance of methodologies available (e.g., confidence intervals, confidence curves, likelihood ratios, severity curves).

This was nothing but bluster between irascible alpha males. Fisher was unclear and Pearson was nitpicking. Pearson must have known that Fisher's "certainly true" was a slip of the pen (Fisher should have simply said, 'supported strongly') and that Fisher, like Pearson himself, generally used "hypothesis" only in the sense of null hypothesis (Inman 1994). The term "alternative hypothesis" had only recently been introduced (Neyman & Pearson 1933b), and was never accepted by Fisher — though he certainly accepted the hypothesis that $\mu_1 \neq \mu_2$ every time he rejected the hypothesis that $\mu_1 = \mu_2$!

Modern reverberations and missteps

Fisher (1955a) waited twenty years to expand

Table 1. Test of three null hypotheses concerning the difference between two treatments.

Treatment group	Values for response variable	*m*	*s*
A	13, 37, 23	24.33	12.06
B	19, 30, 41	30.00	11.00
Possible statistical tests			
Test 1:	H_0: $\mu_A = \mu_B$, $t = 0.605$,	d.f. = 4,	$H_1 = \mu_A \neq \mu_B$ **$P = 0.58$**
Test 2:	H_0: $\mu_A = \mu_B + 1$, $t = 0.498$,	d.f. = 4,	$H_1 = \mu_A \neq \mu_B + 1$ **$P = 0.64$**
Test 3:	H_0: $\mu_A = \mu_B - 1$, $t = 0.712$,	d.f. = 4,	$H_1 = \mu_A \neq \mu_B - 1$ **$P = 0.52$**

on and clarify his point, concluding "it is a fallacy, so well known as to be a *standard* [Fisher's emphasis] example, to conclude from a test of significance that [if *P* is high] the null hypothesis is thereby established …." Neyman (1956) attempted to rebut Fisher but mostly talked around the main point and concluded by pointing out that, well, Fisher himself had once committed the fallacy in the past!

While any good modern textbook warns against this fallacy, in other places confusion still reigns. Ziliak and McCloskey (2008: 69, 225) gratuitously attribute Fisher's (1955a) remarks to his "studied ignorance of Type II error" and to "more despair … by a man losing status with the highbrows." Throughout their book they make the assumption that significance tests are only used in the Neyman-Pearsonian manner where a high *P* value leads to 'acceptance' of the null or at least a favoring of it over the alternative hypothesis. That is why they accuse 94% of the 369 research articles published during 1980–1999 in the *American Economic Review* of commiting the "error" of not giving explicit attention to the power of the tests carried out, and 70% of the "error" of not "examin[ing] the power function" (p. 81, 83). This left those articles *supposedly* "risking high levels of Type II error." Yet that risk of Type II error has been and is non-existent for paleo- and neoFisherians.

In his otherwise fine text, Spanos (1999: 690) briefly becomes very non-Fisherian when he switches from characterizing *P* as a measure of the plausibility of H_0 to suggesting Fisher believed that a *P* value can also provide a measure of "the strength of evidence for … the null hypothesis." Though in most of his text, Spanos seems to accept a neoFisherian paradigm, he gives a strange small table intended to demonstrate a nuanced, non-dichotomous interpretation of the *P* scale. It is as follows: "$P > 0.10$ … strong support for H_0; $0.05 < P < 0.10$ … some support for H_0; $0.02 < P < 0.05$ … lack of support for H_0; $P < 0.01$ … strong lack of support for H_0." Under neither the paleo- nor the neoFisherian paradigm is it logically possible for a *P* value to provide support, let alone "strong support," for H_0.

Spanos makes the same mistake in the example he presents: an observed proportion, 0.48415, is tested for departure from H_0: $\Theta = 0.4857$. This yields a $P = 0.617$ leading him to conclude that "the evidence is strongly in favor of H_0." Yet the implicit H_1: $\Theta \neq 0.4857$ also is highly plausible given this *P* value. And a point H_1: $\Theta = 0.4849$ would yield a $P > 0.617$ and be seen as even more plausible than 0.4875 if it (0.4849) were to be set up as the null. (The small deviations involved in this example should not distract the reader; the same principle is demonstrable with larger ones). A. Spanos (pers. comm.) has indicated that the small table will be removed and these paradoxes resolved in the upcoming new edition of his book.

As many have acknowledged, acceptance of null hypotheses has often been driven by widespread misdefinition of *P*, the mistaken notion that it gives the probability that the null hypothesis is true — or something like that. A current glaring example is provided as an online "public service" by StatSoft (2007; Hill & Lewicki 2007): "The statistical significance of a result is the probability that the observed relationship (e.g., between variables) or a difference (e.g., between means) in a sample occurred by pure chance ('luck of the draw'), and that in the population from which the sample was drawn, no such relationship or differences exist …. Specifically, the *p*-value represents the probability of error that is involved in accepting our observed result as valid, that is, as 'representative of the population'. For example, a *p*-value of .05 (i.e., 1/20) indicates that there is a 5% probability that the relation between the variables found in our sample is a 'fluke'."

But it gets worse. StatSoft (2007) boasts on their website (http://www.statsoft.com/textbook/stathome.html) that their online manual "is the only internet resource on statistics recommended by Encyclopedia Brittanica." Never has it been so important to 'Distrust Authority,' as the bumper sticker says.

The missing example

A high *P* value is as consistent with H_1 as with H_0 and is grounds only for indecision or suspension of judgment with respect to the truth of H_0 (Fisher 1925, Tukey 1960, 1991, Kalbfleisch & Sprott 1976, Oakes 1986: 31, Abelson 1995, Cortina & Dunlap 1997, Harris 1997a, 1997b, Nickerson

2000, Tryon 2001, Lombardi & Hurlbert 2009). One is free to 'accept' H_0 on grounds other than the test and resultant *P* value, but the high *P* value itself provides no grounds for preferring H_0 over H_1. Likewise, it is equally inappropriate to speak of a high *P* value as an argument for 'retaining' or 'failing to reject' H_0 *and*, implicitly or explicitly, 'rejecting' H_1. Both H_0 and H_1 must be viewed as being consistent with the data, and so *both* must be "retained," *both* we must "fail to reject."

An example may help. Table 1 presents a simple hypothetical data set for comparing two treatments. A standard *t*-test of H_0: $\mu_A = \mu_B$ yields a high *P* value (Test 1). Unthinking followers of the decision theoretic framework will accept, on this ground, the null hypothesis and conclude there is no difference between treatments.

That conclusion represents a logical error known as the fallacy of affirming the consequent as proof of the antecedent (e.g. Copi 1953: 251, Henkel 1976: 35, Ford 2000: 177): if A implies B, and B is observed, then A *is* — even though C, D, and E could also imply or explain B. In our example, A = no difference between treatments, B = high *P* value (usually), C = small effect size, D = high variances, and E = low treatment replication.

The fallacy becomes obvious if we carry out *t*-tests for the null hypotheses that μ_A is a little greater (Test 2) and a little smaller (Test 3) than μ_B. Both of these tests also yield high *P* values (Table 1). So, if a high *P* value constituted evidence in favor of the null hypothesis, we would now have evidence that simultaneously favors three mutually contradictory conclusions regarding the likely difference between μ_A and μ_B.

It is unfortunate that such examples are not given in textbooks as soon as the concept of significance testing is introduced. We know of no textbook that does so. On the other hand, the number of books (e.g., Marks 1982: 125, Oakes 1986: 11 (contra p. 31!), Sokal & Rohlf 1995: 65, Underwood 1997: 17, Lang & Secic 1997: 65, Steel *et al*. 1997: 94, Gotelli & Ellison 2004: 96; Hawkins 2005: 84ff) and review articles (e.g., Rozeboom 1960, Morrison & Henkel 1970: 309, Elenbaas *et al*. 1983, Leventhal & Huynh 1996) that promulgate the fallacy is large. Frick (1995) noted four other statistics textbooks, out of 15 examined, that promote it as well.

True parsimony

A particular notion of parsimony combined with the notion that one should not suspend judgment and retain both H_0 and H_1 may sometimes underlie the confusion. For example, Gotelli and Ellison (2004: 91, 92, 96) cogently state that "absence of evidence is not evidence of absence; failure to reject a null hypothesis is not equivalent to accepting a null hypothesis (although it is often treated that way)." However, in discussing an example, they later state, "scientists favor parsimonious or simple explanations over more complex ones. … [and] that differences in GC [glucocorticoid] levels between the two groups can be most parsimoniously attributed to random variation among individuals." In other words, on the grounds of parsimony, they accept or retain only H_0 and presumably would do so for any other statistical test yielding a high *P*. Yet no decision between H_0 and H_1 is required or can be logically justified. So true parsimony calls for retention of both H_0 and H_1.

Logic as "nihilism"

Perhaps thinking of the null hypothesis as a scientific hypothesis rather than just a part of the mechanics involved in obtaining a *P* value, some researchers feel it desirable to expand our conceptual frameworks so that acceptance of H_0 can be a logically valid option, even without recurrence to information external to the data set being analyzed. Indeed, Neyman and Pearson (1933b:187, 195) early on expressed this impulse when they mused about establishing three options — accept H_0, reject H_0, remain in doubt — in their framework, and spoke of "problems … where [type II] errors can be divided into two classes — those which do not matter and those which do …."

Among recent writers, Frick (1995) argues this position most strongly and explicitly. He labels as "nihilism" the idea that a statistical test can never lead to the favoring of H_0 over H_1. He posits that if, by six criteria he advocates, a study constitutes a "good effort" to detect an effect and comes up with a high *P* value, the conclusion should be that the effect is zero. In particular, he

hopes that "Perhaps psychology will eventually settle on one particular number [i.e. *P* value] that must be exceeded for the null hypothesis to be appropriately accepted. While a *p* value of less than .20 seems too low, a *p* value greater than 0.50 seems large enough." Just as he thinks α should be fixed at 0.05 for the field of psychology (Frick 1996 and above), so he also would recommend a critical *P* value to be designated for allowing acceptance of null hypotheses in that field.

The argument will not convince many. It is, at least in part, a reaction to a perceived incorrigibility of his outlaw colleagues. Frick (1995) despairs that "On a practical level, never accepting the null hypothesis is not a viable alternative. Although currently the opinion expressed most often is that the null hypothesis should never be accepted, the null hypothesis is, nonetheless, often accepted. Thus, the nihilistic position seems to be unenforceable." In fact, good scientists and statisticians for the better part of a century have found it quite "practical" to never accept null hypotheses.

Three-decision procedure

We are left, then, with what has been termed the "three-decision problem" (Lehmann 1950), "three-decision procedure" (Kaiser 1960) or "three-valued logic" (Harris 1997a). Its earliest explicit formulation was given by Bahadur (1952) who noted approvingly that "the manner in which the two-sided *t*-test is widely used in practice" entails either concluding there is an effect, with a particular sign, or "reserving judgment." Cox (1958) also gave an early formulation of the idea, stating that "the significance test is concerned whether we can, from the data under analysis, claim a difference in the same direction as that observed … [or] whether the direction of any effects has been reasonably well established …." The idea has been refined and recommended by Kaiser (1960), Tukey (1991), Abelson (1995), Harris (1997a, 1997b), Tryon (2001), and Cox (2006a).

In a two-group case, one examines the *P* value yielded by a significance test for a difference between groups and concludes one of three things: the difference between the true population means seems to be negative, it seems to be positive, or it cannot confidently be stated to be either so judgment is reserved or suspended. We add here only one element — that the interpretation should be a shaded one made without reference to a specified α and without use of terms such as 'significant' and 'non-significant'.

Hunter (1997) strongly disparaged this "three-valued" logic paradigm. He stated: "Harris [1997b] argues that the computations of the significance tests can be saved by using a radically different interpretation scheme. However, his scheme was put forward 35 years ago by Kaiser and was adopted by no one. So even though his new scheme would be an improvement, we already know that it will not work in practice." Kirk (2007) likewise claimed "the three-outcome test [has] … found little acceptance among researchers."

These statement seem naive. Harris (1997a) is certainly correct when he states, "To their credit, most researchers and textbook authors actually follow three-valued logic." This indeed is the logic fully implicit in Fisher's original conception of null hypothesis testing and explicit in writings of early neoFisherians such as Bahadur (1952) and Cox (1958). It seems to have been the quiet guide of *all* those who have used significance tests with understanding of their actual properties and limitations. That is why Harris (1997b) can say, with full justification, that "Adopting three-valued hypothesis-testing logic would require no changes in the conduct of scientifically appropriate research, but only changes in the way we describe the underlying logic of NHST [null hypothesis significance testing] in textbooks, to our colleagues, and to ourselves."

Three-valued logic is thus an important element in the neoFisherian paradigm, a happily positive inheritance from its paleoFisherian progenitor. We modify it only by not requiring specification of an α, thus making the trichotomy a fuzzy one rather than a rigid one.

Hybrids and reformulations

There have been several attempts to ajudicate the historic conflict between the Fisherian and

Neyman-Pearsonian theories of statistical tests. Three are worth comment here. They all move in the direction of our conception of the neoFisherian paradigm without quite getting there.

Lehmann's unified theory

E. L. Lehmann (1993), one of Neyman's early doctoral students and author of the definitive treatise on Neyman-Pearson methods (Lehmann 1959, Lehmann & Romano 2005), suggests "the two theories are complementary rather than contradictory and that a unified approach is possible that combines the best features of both." Which might be more appropriate will vary according to the context or frame of reference of a study. Lehmann suggests "one should routinely report the p value and, where desired, combine this with a statement on significance at any stated [α] level." The later requirement would seem, in any circumstance, to be superfluous. Specification of α is said to be sometimes needed because "definite decisions … are often required" or because "some statisticians (and journal editors) see an advantage in standardization." No examples from research contexts are given where "definite decisions" would be needed. Neither reason seems compelling, nor does Lehmann explicitly indicate how values of $P \geq \alpha$ would be interpreted under his unified theory.

Chow's defense of the hybrid

Chow (1996) gives a strong defense of significance tests and P values. He considers at length the criticisms leveled against significance tests, shows that many are irrelevant or in error, and additionally offers an extended critique of Bayesian methods. Chow may be unique, however, in that the paradigm he chooses to defend is specifically the paleoFisherian-Neyman-Pearsonian hybrid. In response to Gigerenzer's (1993) claim that NHST is "burdened with conceptual confusion," Chow responds (p. 24) that "the hybridism is not necessarily problematic. It is detrimental to research rigour only if it is established that either the [paleo]Fisherian or the Neyman-Pearson approach has to be adopted in its entirety. Moreover, it has to be assumed that either the Fisherian or the Neyman-Pearson treatment is adequate by itself for the task."

Thus Chow (1996) advocates fixing α, talking about 'significant' differences, and 'accepting H_0'. But he is also comfortable with simultaneously reporting exact P values or even using notation such as '< 0.05,' '< 0.01', and '< 0.001', knowing some will interpret such notation to imply having established *a priori* α values of 0.05, 0.01, or 0.001. Chow's objections (p. 38) to the neoFisherian paradigm would seem to be weak: α must be fixed, preferably at 0.05, because that is at least fuzzily "meaningful and objective at the mathematical level" and it gives us an unambiguous "decision criterion." Use of the latter criterion might save some editors and readers some strain on the brain, as discussed, but otherwise has no redeeming value. In sum, Chow's arguments on this point seem to be primarily a rhetorical exercise designed to justify continued use of the word 'significant'.

Mayo's reformulation and severity paradigm

In a wide-ranging and primarily philosophical treatise, Mayo (1996) synthesized a project she began in the 1980s to develop more "rational methods of hypothesis appraisal, … more adequate methods of inductive inference" (p. ix). She rejects "the global inductive approaches … so attractive to philosophers" and comes up with "a model of experimental [i.e. empirical] learning that is more of a piecemeal approach, whereby one question may be asked at a time in carrying out, modeling, and interpreting experiments [i.e. empirical research generally] …" (p. xi). At the core of this paradigm is "a reformulation of standard Neyman-Pearson statistics that avoid[s] the common misinterpretations and seem[s] to reflect the way those methods are used in practice" (p. x). This reformulation she labels *error-probability statistics* or just *error statistics*. Mayo (1997, 2005), Mayo and Cox (2006), Mayo and Spanos (2006, 2009) give briefer, up-to-date summaries of her ideas.

At the core of Mayo's philosophy and *error statistics* is the notion that before a hypothesis,

statistical or scientific, can be accepted it must be subjected to, and pass, a *severe* test. In the context of significance tests, the idea is that *severity* of an inference concerning H_1 can be quantified as $(1 - P)$, the P value being that from a one-tailed test. Severity also can be calculated via additional one-tailed tests for a series of hypothetical discrepancies or effect sizes, that is for inferences more specific than $\partial \neq 0$, such as H_1: $\partial < g$ (which would be tested against H_0: $\partial \geq g$) or H_1: $\partial > g$ (which would be tested against H_0: $\partial \leq g$). One can specify an α in order to have a criterion for which inferences do or don't pass with 'high severity' defined as $1 - \alpha$. One can also calculate and plot a severity curve (g *versus* $1 - P$), without defining an arbitrary boundary between 'high' and 'low' severity, just as a good neoFisherian abandons the arbitrary distinction between 'significant' and 'non-significant' differences, can use a confidence band instead of a confidence interval, and can do power analyses without depending on a single fixed *alpha*. Mayo considers severity analyses to be indispensable to preventing fallacious interpretations of P values, especially low ones.

We attempt no full elucidation of severity analysis here, which would require graphical representations and detailed examples. Exactly in what types of situations such additional apparatus might be both needed and feasible will be a subjective matter. Severity analysis can help discourage fallacious interpretations of P values, although, in our own experience, it is not essential to that end. Manuscripts presenting large numbers of significance tests will be greatly lengthened if a severity analysis is presented for each one. If severity analyses are not essential to a researcher's objectives, they may be viewed with skepticism by reviewers and editors.

In his review of her book, Lehmann (1987) noted that Mayo's views seemed closer to those of Fisher than those of Neyman and Pearson. Though much of her recent work has been in collaboration with apparent neoFisherians such as D. R. Cox and A. Spanos, even her recent works (e.g. Mayo 2005, Mayo & Spanos 2006) use much of the terminology, sometimes with redefinitions, of the Neyman-Pearsonian framework, and she continues to view her work as a "reformulation" or extension of the Neyman-Pearson paradigm rather than a replacement of that paradigm with a neoFisherian one. Mayo patiently critiqued error-ridden early versions of this section, and somewhat agrees (pers. comm.) with the neoFisherian paradigm as far as we have taken it, but she also regards our version of the paradigm as minimalist and incomplete.

Misdirected critiques

Statistical methodologies have been misinterpreted and misused on a large scale throughout their history. Not surprisingly this has stimulated critiques of statistical malpractice on many issues and in many disciplines. Within this corpus of statistical criticism, however, there has been much chastisement of scientists and statisticians for statistical crimes not committed, and much repetition of old erroneous claims. Indeed, the error rate in the literature of statistical criticism may well be as high as that in the disciplinary literatures that the critiques review. In the present context, this applies to many of the claims that P values are used too much and effect sizes and confidence intervals too little. It also applies to attacks by Bayesians who, oblivious to their main intended use, argue for getting rid of significance tests altogether. Arguments on these matters have affected the issue of one-tailed tests mostly indirectly, by creating a great deal of 'noise' in the larger debating arena.

Overemphasis on *P* values?

Most of the time investigators know or suspect on both evidentiary and logical grounds that a Type I error is not likely: there are strong *a priori* grounds for believing that $\partial \neq 0$ (e.g. Savage 1957, Nunnally 1960, Tukey 1960, 1991, Smith 1962, Meehl 1967, Carver 1978, Oakes 1986:39, Cohen 1990, 1994, Chow 1996, Johnson 1999, Stoehr 1999, Anderson *et al*. 2000, Quinn & Keough 2002: 53). Our main focus then should be on description or estimation rather than on hypothesis-testing, e.g., on estimating ∂ (or other measures of effect size) with sufficient precision and not merely on testing whether $\partial = 0$. Yates (1951) long ago noted that "the empha-

sis given to formal tests of significance throughout [Fishers's] *Statistical Methods*, and to a great extent also in The *Design of Experiments* … has caused scientific research workers to pay undue attention to the results of the tests of significance … and too little to the magnitude of the effects they are investigating." This situation persists in many disciplines half a century later and has been widely decried, most recently and vehemently — and confusingly — by Ziliak and McCloskey (2008). The confusions, inaccuracies and polemics in that book distract from its main message, however; these have been partially catalogued by Hoover and Siegler (2008) and Spanos (2008).

This situation in no way constitutes an argument against testing null hypotheses of the form H_0: $\partial = 0$. As Quinn and Keough (2002) nicely state, in the great majority of situations, a "rejection of the H_0 is not important because we thought the H_0 might be true. It is important because it indicates that we have detected an effect worth reporting and investigating further." Others have emphasized the same point (Chow 1996, Eberhardt 2003, Stephens *et al*. 2005). Those many who have disparaged significance testing on the ground that the standard 'nil' null is "silly," "meaningless," "vacuous," "implausible," "obviously false," or "nonsensical" are only reflecting profound confusion about the function of significance tests and the H_0.

Complaining of an overemphasis on *P* values by researchers, many critics (e.g. Rozeboom 1960, 1997, Carver 1978, 1993, Guttman 1985, Oakes 1986, Shaver 1993, Cohen 1994, Falk & Greenbaum 1995, Schmidt 1996, Hunter 1997, Schmidt & Hunter 1997, Royall 1997, Nix & Barnette 1998, Johnson 1999, Anderson *et al*. 2000, Nicholls 2001, Kline 2004, Fidler *et al*. 2006, Wagenmakers 2007, Lukacs *et al*. 2007, McCarthy 2007, Ziliak & McCloskey 2008, Cumming & Fidler 2009) have recommended doing away with significance testing and reporting of *P* values altogether or at least in most situations. We even find occasional premature 'obituaries' that claim "the traditional null-hypothesis procedure has already been superseded in modern statistical theory by a variety of more satisfactory inferential techniques" (Rozeboom 1960) or that make reference to "the collapse of null hypothesis significance testing as a statistical paradigm" (Guthery *et al*. 2001). With few exceptions, however, the complaints of such critics concern the misuse and misinterpretation of significance tests and *P* values by investigators and not the inherent properties, under the neoFisherian paradigm, of the tests or *P* values themselves. Individual scientists and statisticians have been at fault, not the methodology. The critiques are themselves often full of exaggerations and logical and factual errors, as has been abundantly documented (e.g. McGinnis 1958, Abelson 1995, 1997, Chow 1996, Mayo 1996, Cortina & Dunlap 1997, Estes 1997, Hagen 1997, 1998, Mulaik *et al*. 1997, Reichardt & Golub 1997, Rossi 1997, Levin 1998b, MacLean & Ernest 1998, Nickerson 2000, Balluerka *et al*. 2005, Harris 2005, Boruch 2007, Hoover & Siegler 2008, Spanos 2008).

The strong attack by Ziliak and McCloskey (2008: 2) on *P* values and signficance tests starts off on an even keel with words of wisdom: "statistical significance, or lack of it … is on its own almost valueless … [it] should be a tiny part of an inquiry concerned with the size and importance of relationships." Who could disagree? In a report on an aquatic microcosm experiment, Greenwald and Hurlbert (1993) dedicate 23% of the article to graphical representation of effect sizes, report 303 *P* values in the graphs themselves (thus taking up no extra space), and present and discuss the results with no mention of 'statistical significance'. Or consider Hart *et al*. (1998) where 22% of the paper is given over to graphical representations of effect sizes, *P* values in abundance are unobtrusively included within the graphs, and there is nary a mention of 'statistical significance."

Unfortunately, Ziliak and McCloskey quickly get carried away with their own wittiness, neologisms, and iconoclastic fever, producing overblown rhetoric and many inaccuracies about history and statistics. "Statistical significance" is characterized as an "error" (p. xvi) and "a philosophy of mere existence" (p. 7). It is "a view from nowhere … about precisely nothing" (p. 9), and "always a false start" (p. 15), a "mutation" (p. 22), and a sign of "phoniness" (p. 25). A significance test procedure that does not allow acceptance of the null hypothesis and that is

unaccompanied by a power analysis is said to be "meaningless, no better than a table of random numbers" (p. 9). And all this before Chapter 2! Later on, their rhetorical momentum drives them to claim that "Statistical significance is hurting people, indeed killing them" (p. 186). Ziliak and McCloskey simply cannot envisage the possibility of a simple neoFisherian approach like that reflected in Greenwald and Hurlbert (1993), Hart *et al*. (1998) or other papers cited earlier.

Hagen (1997) has perhaps put it best: "The logic of the NHST [null hypothesis significance test] is elegant, extraordinarily creative, and deeply embedded in our methods of statistical inference. It is unlikely that we will ever be able to divorce ourselves from that logic even if someday we decide that we want to. … the NHST has been misintepreted and misused for decades. This is our fault, not the fault of the NHST."

We presume, of course, that the "logic" Hagen refers to is that of the neoFisherian paradigm!

Neglect of effect sizes?

Some modern critics, echoing Yates (1951), have charged that excessive focus on *P* values and "statistical significance" has caused many researchers to even forego presentation of means or effect sizes. Guttman (1985) claimed that "[i] t is rather typical [of the use] of analysis of variance that … the actual means are not published." Of 163 articles in the *Journal of Consulting and Clinical Psychology*, Dar *et al*. (1994) reported that a few reported "differences in proportions or means" but that "no measures of effect size (i.e., eta or omega squared or other measures of the percent of variance accounted for by the independent variables) were ever reported in the context of an ANOVA." With editorial 'encouragement' (Kendall 1997), the percentage of articles in that journal reporting standardized effect sizes rose to 20 by 1993 and to 46 by 2000–2001 (Fidler *et al*. 2005a). On the other hand, of 704 articles published between 1982 and 2000 in *American Journal of Public Health* or *Epidemiology*, essentially all reported effect sizes, yet none used "effect sizes in standard-deviation units" (Fidler *et al*. 2004a). Vacha-Haase *et al*. (2000) determined that only 49% of 1995–1997 articles using significance tests in *Psychology and Aging* and *Journal of Counseling Psychology* reported effect sizes. They also summarized results of nine other studies of psychology journals that found frequency of effect size reporting in the 1990s to range from 10% to 88%. For 95 articles in the *Journal of Wildlife Management*, Anderson *et al*. (2000) stated that "Approximately 47% … of the *P*-values … appeared alone, without estimated means, differences, effect sizes, or associated measures of precision." Graham and Edwards (2001) examined 184 articles that used ANOVA in four major ecological journals published in 1998. They claimed that only two of these articles provided information on effect sizes and that "[w]hen using ANOVA to interpret results of ecological experiments, most ecologists have simply reported *P* values as evidence of, or lack thereof, the biological importance of some factor...on a response variable …." However we examined just the first 10 papers in their set of 184 (those in the first 1998 issue of *Ecology*) and found that at least 9 presented clear information on effect sizes in their tables and figures, most often via group or treatment means. When simple linear regression was also used in some of these papers, slopes of regression lines were given. Expressing how much change in Y is produced or expected per unit change in X, those slopes are effect size measures of primary interest to ecologists.

The real complaint of most of these critics is that many, perhaps most, scientists usually do not calculate and present the standardized measures of effect size so popular with power and meta-analysis *aficionados* over the last decades. These measures include those such as r^2, d/s, ω^2, η^2 and so on, which represent the predictive value of an independent variable or the percentage of total variance explained by it in the particular study. With the exception of r^2 these measures are indeed usually left uncalculated by researchers, because they rarely are useful to clear interpretation of direct measures of effect size (e.g. d, percent change, slope of a regression line, etc.) and the phenomena they bear on. A major value of standardized effect size measures is often said to be their indispensability for

meta-analyses. Yet routine use of measures of predictive value as measures of effect size has massively compromised the meta-analysis literature as many have pointed out (e.g. Oakes 1986: 49ff, 156ff, Hurlbert 1994, Abelson 1995, 1997, Osenberg *et al*. 1997, 1999, Petraitis 1998). Fortunately, warnings of their inappropriateness and potential for misuse are on the increase (Greenland 1998, Levin 1998a, Wilkinson & TFSI 1999, Lenth 2001, Jacard & Guilano-Ramos 2002, Fidler 2002, Di Stefano 2004, Rutledge & Loh 2004, Balluerka *et al*. 2005, Nakagawa & Cuthill 2007, and R. J. Harris, G. Loftus, K. R. Rothman & P. E. Shrout — all pers. comm. to Fidler 2002). Unfortunately, the APA (2001: 25) publication manual and some recent authors (e.g. Kline 2004, Levine *et al*. 2008b) imply that only standardized measures of effect size are legitimate or useful.

Response variables are sometimes defined and measured on artificial or abstract scales. In some disciplines such as animal behavior, psychology and education, this is very common. Effect sizes expressed as absolute increments or decrements or as percent change along such scales can be difficult to interpret and relate to the phenomena under study. Nevertheless, even in such situations conversion of those simple types of effect sizes to standardized ones makes their interpretation less clear, not more clear, even if such conversion can confer a superficial mathematical elegance on poorly conceived and conducted meta-analyses.

Ziliak and McCloskey (2008) argue that Fisher drove many sciences to such an obsessive focus on statistical significance that scientists on a massive scale gave up considering the size and importance of effects in their research reports, costing us "jobs, justice, and lives." Throughout their book they rail against "sizeless scientists", "sizeless sciences", and the "sizeless stare of significance tests." They refer to some of the articles we cite above, and they do analyze in detail several case studies where indeed an obsession with *P* values produced misguided statistical analyses and naïve and damaging interpretation. However, in their own scrutiny of 369 papers in the *American Economic Review* (1980–1999) they do not claim to have found a single paper that omitted information on effect size, much the same result as obtained by Fidler *et al*. (2004a). They do complain that ca. 20%–40% of the papers do not "discuss" or "interpret" the effect sizes they document.

Thus, we consider the allegations of massive underreporting of effect sizes in the natural, behavioral and social sciences to be unfounded. If indeed the matter were more than a 'tempest in a teapot', then it *would* signal a massive, collective, intellectual failure on the part of editors and referees as well as researchers. Consider a simple experiment involving a control treatment and an experimental one. It seems difficult to imagine that a serious researcher would ever prepare, or an editor ever accept, a report on that study that did not give, in tables, figures or text, the relevant response variable means for the two treatments. If that is done, rarely will any standardized effect size calculations be useful, though other supplementary statistics may be, such as confidence intervals, expression of effect as percent change, and so on.

Whether the *importance* or *substantive significance* of effect sizes is sufficiently discussed in a paper is a separate matter and one involving more subjective judgment. Lack of careful interpretation can be a serious problem, as Zilak and McCloskey (2008), among others, emphasize. They suggest that ca. 20%–40% of the *American Economic Review* papers they examined do not "discuss" or "interpret" their effect sizes at all. They suggest the problem is equally severe in other disciplines, but quantitative data on that point are lacking.

More controversial is the claim by Ziliak and McCloskey (2008) that it was Fisher who encouraged everyone to focus on *P* values and forget effect size and importance. They do not make their case. Fisher presented information on effect size in every paper he ever wrote that contained real or hypothetical data sets. Ziliak and McCloskey's complaint that "he rarely *mentioned* [our emphasis] magnitudes — such as the size and meaning of correlation coefficients" (p. 224) — is irrelevant. And their complaint that Fisher recommended we "should not care about the size of experimentally determined death rates or the loss of life in treatment and control groups" (p. 225) is simply vicious and unsupported.

If Fisher did not assess the 'importance' of effect sizes in the data sets he dealt with, surely it was for the same reason writers of statistics books neglect 'importance' in discussing examples they use: it is not appropriate or necessary to their instructional objectives. Cogent assessments of 'importance' are primarily the responsibility of the subject matter specialists, not of the statisticians who may advise them. Moreover, even a subject matter specialist will have limited ability — and space! — to spell out all the many ways in which a given finding may be 'important' to other problems and disciplines or to society at large. Presentation of effect sizes in a clear way that facilitates independent evaluation by others should be the first objective.

Confidence intervals imperative?

Significance tests represent only a small early step in extracting the information in a data set. Various additional information-extracting statistical procedures can be carried out prior to final, more subjective assessment of the import of findings for the phenomena under study. These procedures are so numerous and their utility so context specific, no careful review or comparison of them can be attempted here. Two of the oldest are confidence intervals and power analyses; two of the newest are confidence bands and severity curves. In focusing primarily on confidence intervals, this section is not arguing for their superiority over other procedures but only contesting those who have argued that confidence intervals render significance tests superfluous.

Basic considerations

One common reaction to misuse of *P* values has been the recommendation that instead of presenting them we present the two-sided confidence interval about parameter estimates, as first proposed by Neyman (1937). This of course requires the same largely arbitrary specification of an α as is required in the Neyman-Pearson decision theoretic framework by the test of H_0. Observation of whether the confidence interval does or does not include the value zero (or *c*) does in fact constitute a test of H_0: $\partial = 0$. One does not have to use this overlap criterion in a 'hard' way: degree of acceptance of H_1 can be modulated according to the degree to which the confidence interval overlaps or is distant from zero (e.g. Grant 1962, Cumming & Finch 2005).

So many have been so happy for so long with the 95% confidence interval, we are apt to neglect its arbitrary nature in this situation. There never was, of course, any requirement that the α selected for the test of H_0 by an (unenlightened paleoFisherian or Neyman-Pearsonian!) investigator be the same as the α used in calculating a confidence interval. They could, for example, carry out significance tests with $\alpha = 0.05$ while calculating a 90%, 80% or 75% confidence interval by setting $\alpha = 0.10$, 0.20 or 0.25. We are sympathetic to Cohen's (1990) suggestion that "our interests are often better served by more tolerant 80% intervals." In many situations, especially where temporal trends in a response variable are being shown graphically for two or more groups or treatments, the shorter confidence interval obtained with an $\alpha > 0.05$ will improve figure clarity. This is also why many authors reasonably elect to show standard error bars around their means rather than the traditional 95% confidence intervals.

One valuable function of confidence intervals is the provision of "humility" as they usually are wide (Harris 1997b). This suggests the option of calibration to personal psychological need. During years when things are going badly in the lab, the researcher can seek the solace of 80% confidence intervals. To moderate hubris during good times, 99% confidence intervals could be used.

Emil Spjotvoll (p. 65 in Cox 1977) was one of the first to contest the idea that confidence intervals are more useful and informative than tests. He stated, "My feeling is that they contain different kinds of information. When working with confidence intervals we use a fixed confidence level and hence we do not have the flexibility that [the exact *P*] gives us in measuring inconsistency with a given hypothesis. We could, of course, write up the intervals corresponding to a number of levels or finding [sic] ways of representing this graphically, but I believe this will

probably be more confusing than illuminating." Many others have also opined that confidence intervals complement significance tests and *P* values and cannot serve as replacements for them (e.g. Frick 1995, Cortina & Dunlap 1997, Levin 1998b, McLean & Ernest 1998, Harris 2005). Presentation of multiple confidence intervals for individual estimates does have its champions and might sometimes be useful with very simple data sets. Rozeboom (1960) suggested that reports might with "some benefit … simultaneously present *several* confidence intervals for the parameter being estimated." Salsburg (1985) proposed using 50%, 80%, and 99% confidence intervals simultaneously in clinical studies. Mayo and Cox (2006) state that "the provision of confidence intervals, in principle at a range of probability levels, gives the most productive frequentist analysis."

Appraisal of the value of confidence intervals must also consider that, in most studies, their width will have little bearing on interpretation of the effects demonstrated. So long as the associated *P* value is low, we conclude that an effect of some particular sign and approximate magnitude has been demonstrated. In studies testing broad theory or extraordinary claims (e.g. of telepathic ability), confident knowledge of the existence and sign of an effect is sometimes the sole objective. In those cases, a low *P* value by itself may provide sufficient confirmatory documentation (Frick 1996, Chow 1996, Abelson 1997, Balluerka *et al*. 2005). In most other types of studies in both basic and applied research, the magnitude of the effect is also of direct interest. But only in some fraction of these will routine calculation of confidence intervals be useful.

The width of the confidence interval usually contains no information on the phenomenon being investigated nor does it aid assessment of the external validity of a finding, the degree to which it is generalizable to other and/or larger contexts or systems. The width is an artifact of study design protocols, in particular, in the case of an experiment, the number of experimental units employed and the steps taken to achieve some degree of homogeneity among those units. In some contexts, such as clinical medical research, it is highly desirable to examine variability of response among experimental units (usually patients) in order to understand its possible relation to stratification factors (e.g., age, sex, race) and implications for clinical practice. In such contexts, the more useful sorts of confidence intervals might be ones that express the variability of response, not the precision with which mean response is estimated (Salsburg 1989, 1992, Yancey 1996).

The crusade

The largely self-evident facts summarized above are much at odds with an almost religious crusade that began some decades ago to demand that authors greatly increase the reporting of confidence intervals in research papers and greatly decrease the reporting of *P* values. The fields of medicine and psychology have been the main arena for this debate. Recent papers by F. Fidler and her colleagues (Fidler 2002, Fidler *et al*. 2004a, 2004b, 2005a, 2005b, 2006, Cumming *et al*. 2007) give an excellent overview of the crusade, albeit from the point of view of the current crusade leaders themselves. They also collectively offer a full bibliography on the topic.

Following accumulation of a small arsenal of critiques in the literature, a few editors led the charge. In 1977 under the influence of K. R. Rothman, the *New England Journal of Medicine* began strongly urging more use of confidence intervals and decreased reporting of *P* values. When he became assistant editor of the *American Journal of Public Health* in 1983 and founding editor of *Epidemiology* in 1990, he became more demanding in these matters, sometimes advising authors, "All references to statistical hypothesis testing and statistical significance should be removed from papers. I ask that you delete *p* values as well as comments about statistical significance" (as quoted by Fidler *et al*. 2004b). On the basis of writings by D. G. Altman and S. Gore, *The British Medical Journal* by 1986 was recommending use of confidence intervals in place of *P* values. Perhaps by 1988, the infidels had rallied, and the tide had turned in favor of moderation and balance. By that year, "over 300 medical and biomedical journals had notified the International Committee of Medical Journal Editors (ICMJE) of their willingness to

comply with the [ICMJE] guidelines for publication" (Fidler *et al*. 2004b). The current version of those guidelines states, "When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as use of *P* values, which fails to convey important information about effect size" (ICMJE 2007: 35). Fine advice, but isn't it what good scientists and good statisticians have been doing for more than half a century? In his review of this debate, Walter (1995) noted that authors in epidemiology now appropriately use both techniques and warned about the setting of overly rigid editorial policy on such matters. Though aimed primarily at medical researchers, Altman *et al*. (2000) is an excellent guide to the calculation of confidence intervals for a variety of types of data.

In psychology a somewhat similar trajectory has been followed by editors and editorial boards, and this is well reviewed in Fidler (2002), Fidler *et al*. (2004b, 2005a, 2005b) and Cumming *et al*. (2007). Though the crusade in psychology at the editorial level got a later start than had that in medicine, it has been conducted with greater vehemence. Irresponsible hyperbole such as "the use of statistical significance testing in the analysis of research data has been thoroughly discredited … [it] retards the growth of knowledge; it never makes a positive contribution … instead use confidence intervals and point estimates …" (Schmidt & Hunter 1997) has abounded in the psychological literature for decades. Ecologists, unfortunately, are now being encouraged to join this crusade (Johnson 1999, Di Stefano 2004, Fidler *et al*. 2004b, Fidler *et al*. 2006).

The more aggressive crusaders for "reform" in psychology have advocated doing it by revising the influential APA Publication Manual to make it more rigidly prescriptive on statistical matters. The recommendations of a task force (Wilkinson & TFSI 1999) formed to suggest changes in the manual did not satisfy some reformers (Fidler 2002 and persons quoted therein). Those recommendations did state bluntly and without qualification that, "Interval estimates should be given for any effect sizes involving principal outcomes," but they did not forbid, or even discourage, use of significance tests. The Statistics Task Force that revised the manual itself retained broad-minded, non-dictatorial statements about both significance tests and confidence intervals. The manual (APA 2001: 22) only says the latter "can be an extremely effective way of reporting results … [and are] therefore strongly recommended."

Clutter *versus* clarity

When editors have taken hard stands for or against particular statistical practices, authors not surprisingly have changed their behavior. Recommendations against reporting *P* values and in favor of reporting confidence intervals have had definite, sometimes large effects in the direction desired by editors (e.g. Fidler *et al*. 2004a, 2005a, 2006, Cumming *et al*. 2007). Those changes invariably have been labelled as positive "reform" by the crusaders. Yet no logic can justify tallying as a positive act every use of confidence intervals or tallying as a negative act every use of a significance tests. So before we could label the documented changes in statistical practice a *positive* "reform", the difficult task would have to be carried out of assessing how each statistical decision in a large set of articles contributed to the clarity and cogency with which the substantive findings of the research are presented. No one has done this yet, and probably nobody wants to!

Fidler *et al*.'s (2006) analysis of statistical practices in the journals *Conservation Biology*, *Ecology*, and *Journal of Ecology* is an excellent case in point. In three tables, they present the frequency in these journals of a large number of statistical practices, both 'good' and 'bad'. For each of the 64 frequencies reported, its 95% confidence interval is also reported, an act all *bona fide* crusaders would applaud. Neither individually nor collectively are these confidence intervals interpreted by Fidler *et al*. in any way, however. They serve no function. Their complete excision from the article would require no change in text or conclusions, and would permit a large reduction in sizes of the tables. Yet these authors castigate other researchers for the same failure. After making favorable comment

on increased use of confidence intervals in *Epidemiology* and the *American Journal of Public Health*, Fidler *et al*. (2004a) complained that, "In both journals, however, when CIs were reported, they were rarely used to interpret results or comment on precision. This rather *ominous* [our emphasis] finding holds even for the most recent years we surveyed."

Research projects vary greatly in size, scope and complexity, as do the articles reporting them. In many type of articles, no confidence intervals will be wanted or needed; they were used, for example, in only one of the eleven, effect size-focused, ecological papers cited at the beginning of this article. In other studies, their sparing use will be sufficient, and in yet others they might be useful for every effect size estimated. Where large numbers of frequencies, means, effect sizes, and other statistics are reported, confidence intervals will often add only clutter, just as they do in Fidler *et al*. (2006), not clarity and cogency. Rote use of confidence intervals will thus contravene the wise advice in the APA (2001) publication manual to always seek the "minimally sufficient analysis", as well as the advice of the Ecological Society of America to authors that, "The purpose of statistical analysis is to increase the conciseness, clarity and objectivity with which results are presented and interpreted, and where an analysis does not serve those ends it probably is inappropriate" (ESA 2006).

The more zealous advocates of abandoning explicit significance tests and the reporting of *P* values in favor of reporting of confidence limits exclusively usually have a narrow frame of reference, that of small experiments or studies - those involving few treatments and only one or a few response variables or monitoring dates. In many disciplines large numbers of response variables and monitoring dates can result in tens or hundreds of comparisons among treatments being of interest. While *P* values for all such comparisons usually can be reported without causing large increase in size and complexity of tables, figures, or text, the same would not be true for confidence intervals. Even in small studies where the number of implicit or explicit comparisons is limited, often more useful than the standard confidence intervals will be Tryon's (2001) inferential confidence intervals, which are calculated in such a way that whether or not two intervals overlap corresponds to whether or not a test for a difference between the two means will yield $P > \alpha$ or $P < \alpha$. This procedure is approximately equivalent to calculating a confidence interval about the difference between the means and observing whether it includes zero. When there is need for additional "worrying the bone" of data sets, one can calculate other adjunct measures such as counternulls (Rosenthal & Rubin 1994), confidence distributions (Poole 1987, Schweder & Hjort 2002, Schweder 2003, Bender *et al*. 2005), probabilities of replication (Krueger 2001, Killeen 2005) or severity curves (Mayo 2005, Mayo & Spanos 2006). These can help illuminate the meaning of *P* values thus discouraging their misinterpretation and providing complementary perspectives. Coverage of them even in introductory statistics courses would be salutary even if their actual use will not be needed very often.

Fiona Fidler (pers. comm.) kindly critiqued this manuscript for us and offers that we agree on most major points. We agree that rigid institutionalization or prohibition of any one technique would be counter-productive, that misuse and misinterpretation of significance tests has been the main problem, and that "best statistical practice requires consideration of the full range of possible statistical techniques and researchers' informed judgement to choose the most appropriate design, measure and analyses to serve the particular research goals" (Fidler & Cumming 2008). Her one key disagreement is that she believes significance tests must be deemphasized and used less frequently because our students and colleagues will never learn to use them appropriately. We respond by asking that the neoFisherianism be given a chance. If our students and colleagues have not responded well to force-feeding with the paleoFisherian and Neyman-Pearsonian paradigms, perhaps that speaks well to their intelligence.

Fallacy of the obese *n*

The demon of the overlarge sample: It lurks quietly in the darkness, waiting for researchers

to pass by who are too focused on obtaining adequate sample sizes. If sample sizes are too large, one may be "in danger" of getting very low *P* values and establishing the sign and magnitude of even small effects with too much confidence. Oh, the horror of it all.

Though this "danger" has repeatedly been called to our attention for half a century, few have paid much heed — fortunately. The concern derives from intersection of the generally accepted fact that almost all nil null hypotheses are false with the fact that the primary value of low *P* values is indeed as indicators of confidence in the estimated sign and magnitude of effects. Berkson (1938) was an early instigator of the confusion, saying: "an observant statistician will agree that … when the numbers in the data [i.e. sample sizes] are quite large, the *P*'s tend to come out small …. If this be so, then we have something here that is apt to trouble the conscience of a reflective statistician …. If we know in advance the *P* that will result from an application of a Chi-square test to a large sample, there would seem to be no use in doing it on a smaller one." Carver (1978) discussed how "Controlling experimenter bias is a much discussed problem, but not enough is said about the experimenter's ability to increase the odds of getting statistically significant results simply by increasing the number of subjects in an experiment. In effect, Carver labeled as "bias" any attempt to gain power by increasing sample size. Good (1982) claimed that "a given *P* has dimishing significance as *N* increases" and used a Bayesian rationale for suggesting that a "standardized *P* value" be reported; where $N > 100$, this would approximately equal the *P* value that would have been obtained if N had been exactly 100. Serlin and Lapsley (1985) argued for "adopting a methodology … that, even with infinite sample size, does not always reject the null hypothesis [when it is false]." Anderson (1987) suggested "the appropriate significance level should be adjusted to sample size." Levin (1998b) says that "intelligent hypothesis testing…will be based on sample sizes that are … not so large as to detect effects that are deemed to be substantially trivial." Anderson *et al*. (2000) reason that "using a fixed α-level to decide to reject or not reject the null hypothesis makes little sense as sample size increases … [so] theoretically, α should go to zero as *n* goes to infinity." Daniel (1998) notes that "for a given statistical effect, a large sample is more likely to guarantee the researcher a statistically significant result than a small sample is … [with a large sample] even inordinately trivial differences between the two groups could be statistically significant … [a statistical signficance test] is largely a test of whether or not the sample is large …." Ziliak and McCloskey (2008: 67, 81) accuse 22% of the 1980–1999 research articles in *American Economic Review* of the "error" of failing to "use a small number of observations, such that statistically significant differences are not found merely by choosing a very large sample." Levine *et al*. (2008a) announce that "Perhaps the most widely recognized limitation in NHST is its sensitivity to sample size …. When sample sizes are large, even trivial effects can have impressive-looking *p* values." Rindskoff (1997), Nix and Barnette (1998), Thompson (1998), Marden (2000), and dozens of other writers seem to concur that the overlarge sample is a real "danger".

Proponents of the danger of the obese *n* seem to worry too much about the 'bottom of the class' — those persons who confuse statistical significance with importance or substantive significance, and those who confuse strong confirmation of a prediction with strong confirmation of the theory or scientific hypothesis that generated the prediction. Such proponents want to keep low *P* values out of the hands of the 'bottom of the class' for the same reason we should keep matches out of the hands of children. But outlawing matches, i.e. powerful tests, does not seem a reasonable solution. The fallacy of the obese *n* has been well discussed by Mayo (1985, 1996: 401–403, Mayo & Spanos 2009), who also points out that this fallacy has sometimes been part of the weak arsenal with which Bayesians have attacked frequentist statistics.

Statistical hypotheses *versus* scientific hypotheses

Confusion between scientific inference and statistical inference has abounded in the statistical and disciplinary literature for much of the

past century. It has been most conspicuously reflected in the common failure to distinguish the concepts of *statistical hypotheses* and *scientific hypotheses*. In particular, many authors refer to the alternative statistical hypothesis (H_1) as the "scientific" or "research" hypothesis (e.g. Feinstein 1975, Carver 1978, Gigerenzer & Murray 1987: 12ff, Daniel 1990: 6, Wolterbeek 1994, Johnson 1999, Quinn & Keough 2002: 5). Thus we often find statements like: "the point of statistical analysis in ecological research is the testing of a *scientific* hypothesis that the imposed treatment had the hypothesized effect;" (Ellison 1996); "traditional significance tests present *p*-values as a measure *against* a theory" (Thompson 2006); or "The accepted standard in most of ecology … is that a claim for a successful theory requires rejection of a reasonable null hypothesis" (Gotelli & McGill 2006). Since true scientific hypotheses usually have a 'directional' character, the conflation of the two concepts has, for some scientists, lent an air of legitimacy to directional statistical hypotheses and 1-tailed tests.

Clear distinction despite Fisher

R. A. Fisher has sometimes been blamed for this confusion. Schmidt and Hunter (1997: 42), for example, claim the fact "that many researchers believe that null hypothesis significance testing and hypothesis testing in science in general are one and the same thing is a tribute to the persuasive impact of Fisher's writings …. In his writings, Fisher equated null hypothesis significance testing with scientific hypothesis testing." But did he? There is no hard evidence that Fisher failed to appreciate the distinction between statistical hypotheses and scientific hypotheses. However, his frequent lack of clarity and the particular titles he chose for some works clearly misled people. Among those titles were: *The logic of inductive inference* (1935c), *Statistical methods and scientific induction* (1955), and *Statistical methods and scientific inference* (1956). He tended to get into trouble when he moved from the realm of mathematical statistics, where he was king, into the realm of the philosophy of science, where he was not and where precision of symbolic notation less easily offsets verbal imprecision. As Kendall (1963) noted, "Fisher had no gifts of exposition, even of his own ideas, and rarely set out explicitly the assumptions on which he was working." Fisher was "a genius of the first rank, perhaps the most original mathematical scientist of the century. A difficult genius though, one in whom brilliance usually outdistances clarity" (B. Efron, in discussion of Savage 1976).

Fisher's foibles are a weak excuse, however, for us to invoke in the 21st century and decades after the distinction between scientific and statistical hypotheses has been repeatedly clarified. Cox (1958) emphasized the distinction between "statistical inference" and "scientific inference" and how small a role the former sometimes plays in the latter. Anscombe (1961) stated "All scientific theories ultimately rest on a simple test of conformity: universal hypotheses are confirmed by noting the incidence of favorable cases, statistical hypotheses are confirmed [or at least supported] by significance tests." Bolles (1962) amplified this, stating, "The final confidence [a scientist] can have in his scientific hypothesis is not dependent upon statistical significance levels; it is ultimately determined by his ability to reject alternatives …. The effect of any single experimental verification [of a prediction] is not to confirm a scientific hypothesis but only to make its *a posteriori* probability a little higher than its *a priori* probability …. One of the chief differences between the hypotheses of the statistician and those of the scientist is that when the statistician has rejected the null hypothesis, his job is virtually finished. The scientist, however, has only just begun his task." Clark (1963) succinctly said, "Statistical hypotheses concern the behavior of observed random variables, whereas scientific hypotheses treat the phenomena of nature and man." Meehl (1967) said, "It is important to keep clear the distinction between the *substantive theory* of interest and the *statistical hypothesis* which is derived from it …. [I]n practice there is a tendency to conflate the substantive theory with the statistical hypothesis, thereby conferring upon [the substantive theory] somewhat the same degree of support given H [H_1] by successful refutation of the null hypothesis." Henkel (1976: 34) noted, "A statistical hypothesis is a statement about a

population parameter, or parameters. Statistical hypotheses are usually not the same as the substantive, or scientific, hypotheses that we wish to test, but should be a logical consequence of the substantive hypotheses." Finally, we can cite Simberloff (1990): "Scientific hypotheses are about phenomena in nature. Statistical hypotheses are about properties of populations based on samples …. Rejection of one or more statistical hypotheses would constitute one piece of evidence to be weighted in deciding whether to reject a scientific hypothesis." There would not seem to be much room for confusion. But for some people the force or *cachet* of 'hypothesis' may simply overwhelm the distinction attempted by the qualifiers 'scientific' and 'statistical'.

In most contexts, statistical hypotheses are only of two sorts: null (H_0) and alternative (H_1). There are various ways of structuring these according to where we wish to assign the burden of proof in a test. For example, as discussed earlier, H_0 could be that ∂ is zero, is some particular non-zero value, or is greater or lesser than some particular value. Neyman (1950) preferred the term "hypothesis tested" to "null," to avoid any implication that it was always that $\partial = 0$, but the suggestion was never widely accepted. In sum, statistical hypotheses are simply part of the mechanics of carrying out an assessment or summary of a given data set using simple formal procedures that help determine what the data set may or may not tell us with some degree of certainty. A scientific (or substantive or research) hypothesis, on the other hand, is tested only by assessments of multiple data sets or pieces of information, many of these assessments perhaps involving no statistical tests of any sort at all.

The same distinctions can be made in the language of models without any recourse to the term 'hypothesis'. Thus Spanos (1999: 544) states, "Statistical models are viewed as *first stage models* in the sense that their primary goal is to provide *statistically adequate descriptions* of observable stochastic phenomena; statistical models do not pretend to offer explanation …. [They] are specified exclusively in terms of the observable random variables that presumably have given rise to the data. This should be contrasted with theory models which are defined in terms of theoretical concepts that might or might not have a direct connection with observational data."

Unfortunately these waters are continually being muddied over and over again, especially in the literature of statistical criticism. As a recent example, Lukacs *et al*. (2007) state that "statistical models should represent a translation of scientific hypotheses to their equivalent mathematical expression. … The science hypotheses and statistical models should always be very tightly linked." Failure to keep clear the distinction between statistical hypotheses (or models) and scientific hypotheses has driven many of the attacks on Fisherian significance testing by advocates of Bayesian, likelihood, information theoretic and even Neyman-Pearson methods. For these reasons we have treated the matter here at greater length than its simplicity might otherwise call for.

Same distinction in applied research

It is sometimes implied that, in practical contexts where the scientific question seems simple — will this medicine improve patient survival? will this fertilizer increase yield? will this teaching method improve student test scores? — there is a one-to-one correspondence between the scientific hypothesis and the alternative hypothesis. But this results from a simplistic view of what the "practical scientist" must demonstrate. Chow (1987) is the clearest writer we find on this matter. He contrasts the testing of a fertilizer with the testing of a psychological theory, noting that issues of "generality" are relevant in both contexts and imply the need in each case of *many* studies and perhaps *many* sets of null and alternative hypotheses. In the fertilizer study, implicitly the usual context would entail assessment of the fertilizers effectiveness under different soil and climatic regimes and different crop management practices as Yates (1951) had also pointed out. Otherwise the reliability of advice to farmers may be very doubtful. So a multiplicity of similar experiments might be called for to give them high collective generalizability or external validity. And a professional agronomist, if not his field technicians, may well have interest in further questions and experiments aimed

at elucidating, for example, the physiological mechanisms determining the optimal fertilizer application rate and so on.

For testing a psychological theory and its generalizability, Chow (1987) emphasizes that a theory consists of a set of propositions and thus requires a variety of quite *dissimilar* experiments and other studies for its testing. However, a multiplicity of *similar* experiments may also be called for in order to document the generalizability of the theory. Does it hold for both men and women, for children and adults, for persons in different cultures? The point, then, is that it will be rare that any scientific question of practical or theoretical import will be reducible to a single statistical alternative hypothesis or model.

A common delusion

Some of the confusion beween scientific and statistical hypotheses derives from the delusion, that strong evidence against H_0 represents not only strong evidence in favor of H_1 (which it does), but also strong evidence in favor of the scientific hypothesis or theory that predicted the failure of H_0 — despite all the other scientific hypotheses that might have predicted the same failure of H_0. We have quoted Meehl's (1967) early warning above, and many other attempts have been made to correct that delusion (e.g. Meehl 1997, Henkel 1976, Oakes 1986, Kline 2004, Levine *et al*. 2008a). Though it is an obvious one, it is a psychologically attractive one in that to provide strong confirmation of a theory is clearly more heroic than to simply provide one more data point in its favor, which is usually all that a single statistical test or single study accomplishes.

Some writers have proposed, unfairly we believe, that significance testing itself bears much responsibility for this delusion. Thus, Howard *et al*. (2000) state that "the most serious limitation of NHST is that it has led researchers to focus their efforts on designing a single study to address scientific hypotheses … it is rarely that a single study … can be viewed as providing a definitive test of a scientific hypothesis." Researchers, heal thyselves. No fault lies with the significance test!

Null hypotheses and null models

A final semantic problem is one perhaps particular to ecology ever since debates in the late 1970s and 1980s over problems of community ecology and island biogeography (Strong *et. al*. 1984). It is the confusion of statistical null hypotheses with so-called 'null models'. That has led to statements such as, "null hypotheses are not frequently used in ecology" (Strong 1980: 273, Ford 2000: 217); "Null hypotheses in ecology are often unsatisfactory because they are virtually impossible to specify completely …" (Quinn & Dunham 1983); and "it may not be possible to construct null hypotheses because we cannot specify what may exist in the absence of a particular factor" (Ford 2000: 218). All these statements are intended to refer to null models. In some places Quinn and Dunham *do* put quotation marks about "null hypothesis" when they use it to mean null model, however, and in general they do give a good analysis of the restricted utility of null models in ecological research, as do the sharp debates in the collection edited by Strong *et al*. (1984).

Where a null model is very specific and developed for a very restricted domain, it perhaps can be translated into a single statistical null hypothesis. However, when the null model is a broader statement it may have more the character of a *counter*-research hypothesis, a negation of the influence of a factor, a proposition of "nonexistence of the cause" (Quinn & Dunham 1983). An example would be the proposition that competition does *not* influence the structure of island bird assemblages. In theory one can determine what the structure of the assemblages would be if competition had had no influence on them, and then compare that, perhaps even with a statistical significance test, to the actual observed structure. But the usual impossibility in practice of coming up with a single, cogent, defensible null model usually will mean that a counter-research hypothesis, like its counterpart research hypothesis (e.g. competition *does* influence the structure of island bird assemblages), can only be tested via a multitude of separate studies and analyses of more focused questions, using not only significance tests but also likelihood and information theoretic model selection methods. In any case, confusion

may be avoided if we use null hypothesis in these contexts only with the specific significance tests that may be used in evaluating the predictions of a null model.

Complaints of Bayesians and likelihooders

Up to this point we have considered primarily criticisms of significance tests and *P* values that have come from the frequentist schools of statistics. Broadly construed, these include Neyman-Pearsonians, paleoFisherians and neoFisherians, though some writers use 'frequentist' exclusively for the Neyman-Pearson tradition where α is fixed and *P* values need be reported only as $> \alpha$ or $\leq \alpha$. Their grounds of complaint have been that significance tests are superfluous or that they are excessively used or carelessly misinterpreted and that scientists are incorrigible and should not be allowed 'to play with matches'.

P values exaggerate evidence?

Many champions of Bayesian, likelihood, and information-theoretic methods level harsher charges against significance tests. They claim that they are logically flawed at a fundamental level and do not provide the information that even their most careful users think they do. Perhaps the two harshest specific charges are that *P* values exaggerate the evidence against null hypotheses and that they are based on "unobserved data." Let us look at each of these complaints and clear up the illogicalities and misunderstandings that have generated them. More comprehensive critical assessments of Bayesian statistics can be found in Lecam (1977), Efron (1986), Chow (1996), Spanos (1999), Dennis (1996, 2004), Mayo (1996, 1997) and Cox (2006a, 2006b).

Let us start by presenting the first claim in the words of several of its believers:

"*… classical procedures are often ready severely to reject the null hypothesis on the basis of data that do not greatly detract from its credulity, which dramatically demonstrates the practical difference between Bayesian and classical statistics*." (Edwards *et al*. 1963)

"*… P values can be highly misleading measures of the evidence provided by the data against the null hypothesis … p gives a very misleading impression as to the validity of* H_0 *from almost any evidentiary viewpoint … actual evidence against a null (as measured, say, by posterior probability or comparative likelihood) can differ* by an order of magnitude *from the P value*." (Berger & Sellke 1987)

"*Bayes factors show that P values greatly overstate the evidence against the null hypothesis*." (Goodman 1999b)

"*Small values of P are taken to represent strong evidence that the null hypothesis is false, but workers demonstrated long ago that such is not the case*." (Johnson 1999)

"*… the P value generally overstates the evidence against* H_0*, i.e. it rejects the* H_0 *when the posterior probability suggests that the evidence against* H_0 *is relatively weak*." (Quinn & Keough 2002: 56)

"*Comparison to a very general Bayesian analysis shows that p values overestimate the evidence against the null hypothesis …. The p value does not quantify statistical evidence*." (Wagenmakers 2007)

Similar statements are found in Berger and Delampdy (1987), Falk and Greenbaum (1995), Goodman (2003), Hubbard and Bayari (2003), and throughout the Bayesian literature.

The potency of these claims is sometimes reinforced by misrepresentation of what Bayesian methods themselves can provide. For example, Mills (2003) claims "it is possible to determine the probability of the null hypothesis, given the data at hand if one uses the Bayesian approach to statistical analysis"; Berger (2003) reports that a "conditional frequentist error probability … precisely equal[s] the objective Bayesian [posterior] probabilit[y] and … is the probability that the hypothesis is true …"; McCulloch (2004) states that "Bayesians can calculate the probability that a null hypothesis is false. This has a much more straightforward interpretation than the awkward definition of a frequentist *P*-value and is what many scientists would like to calculate;" Ellison (2004) opines that "Bayesian methods … pro-

vide a direct measure of the probability of one or more hypotheses of interest;" Gigerenzer *et al*. (2004) claim that "Unlike null hypothesis testing, Bayes' rule can actually provide a probability of a hypothesis;" and McCarthy (2007: 52) states that "Perhaps the main defining feature of Bayesian methods is calculation of the probability of a hypothesis being true."

All of these statements about the weak evidential nature of *P* values and about the meaning of Bayesian posterior probabilities are false.

Two Bayesian examples dissected

A full review of the extensive literature on Bayesian *versus* frequentist methodologies is not needed to demonstrate the erroneous nature of the above Bayesian claims. Let us consider just two sets of examples that above authors use to try to make their case. In the interest of brevity, we presume familiarity with basic Bayesian concepts on the part of the reader.

Berger and Sellke (1987) give examples where the point null H_0: $\partial = 0$ is being tested against the standard composite H_1: $\partial \neq 0$. They apply a so-called "objective" or "non-informative" Bayesian approach where H_0 and H_1 are both assigned a prior probability of 0.5. Not surprisingly, data sets yielding a *P* value of 0.05 yield Bayesian posterior probabilities severalfold higher. That is interpreted to mean that, despite the *P* of 0.05, "there is at best very weak evidence against" H_0. They imply that the posterior probability is the true "magnitude of the evidence against H_0."

Bayesian priors can yield results reflecting not just an investigator's true *beliefs* but also political, financial or religious motivations. Such results could damage science or society, at least in the short run, in the hands of statistically unsophisticated decision makers. Dennis (1996, 2004) gives examples of investigators concerned about management of rare species, pollutant concentrations in rivers downstream from mining operations, and efficacy of dietary supplements, and wonders whether such investigators' 'prior beliefs' about those situations might vary according to where their salary or research funds were coming from. In a rather different sphere, Unwin (2003) uses a Bayesian approach to estimate the probability of the existence of god. A good 'objective' Bayesian, he gives 'exists' and 'does not exist' both a prior of 0.50, then evaluates a data set consisting of six facts, and ends up with a posterior probability of 0.67 in favor of 'exists,' which is then upgraded to 0.95 by an additional injection of personal belief. As Dawkins (2006: 132) notes in his critique of this analysis, "It sounds like a joke, but that really is how he [Unwin] proceeds …. I can't get excited about personal opinions, whether Unwin's or mine."

The relative and subjective nature of Bayesian 'probabilities' needs to be fully comprehended, as does the fact that none are probabilities of the truth of hypotheses. Bayesian priors are guesses of the investigator or imposed by some 'objective' convention. They are degrees of belief assigned non-zero values only for the specific point hypotheses or models in the set considered. A point hypothesis or model outside that set may be superior to, or closer to the truth than, any in the set. The priors are labeled as "subjective," "personalistic," "objective," or "reference" "probabilities" of the truth of an hypothesis. Bayesian posterior probabilities are then just guesses modified by additional information. In much of their writings, especially those criticizing significance tests, Bayesians drop all qualifiers and just claim to be estimating the "probability of truth". "In the world of Bayesian statistics, truth is personal and is measured by blending data with personal beliefs" (Dennis 2004).

One of the major logical incongruities here is that significance tests in all disciplines are mostly used where H_0 is known or strongly suspected *a priori* to be false. So by Bayesian logic it should be assigned a low prior probability, e.g. 0.01 or 0.10. Casella and Berger (1987) note this would result in a much lower posterior probability. Berger and Sellke (1987), referring to a hypothetical example, suggest, however, that even using a prior as low as 0.15 for H_0 would constitute "blatant bias toward H_1 [and] … hardly be tolerated in a Bayesian analysis." So much for the desirability of using priors to express prior information or personal belief. The "bias" responsible for low priors and such contradictions is better labeled the *wisdom* of the investigator in selecting for study, independent

variables that indeed do influence or are correlated with the dependent variables of interest. As Royall (1997: 73) has noted, attempts to find completely objective or 'non-informative' priors "have been unsuccessful for a simple reason — pure ignorance cannot be represented by a probability distribution."

Even more damning to Berger and Sellke's claims is the fact that if the prior probability of H_0 is set at < 0.35, and if the observed *P* value is 0.05, then the posterior probability for H_0 will always be < 0.05 (Krueger 2001). It would be rare that a subjective prior for H_0 of > 0.35 could be justified. That prior usually should be much lower as we have suggested. Thus, as compared with Bayesian analyses with low priors for H_0, standard significance tests will generally *underestimate* the evidence against H_0, at least if we assume — as Bayesians often do — that a *P* value and a Bayesian posterior probability can be meaningfully compared simply because both can range from 0 to 1.

For much the same reasons we reject as unhelpful the suggestion (Selke *et al*. 2001, Berger 2003) that the lower bound on the objective posterior probability of H_0 will be of any value as a "quick and dirty calibration" against which to compare *P* values when H_1 is the standard composite alternative one. Critical comments of seven discussants of Berger (2003) printed at the end of that article, as well as Berger's rejoinder, address this and related issues further.

Not willing to yield ground in the face of the arbitrariness inserted into analyses by the subjective nature of all systems of priors, clever Bayesians will find a fallback position. Goodman (1999b), for example, claims that "The minimum Bayes factor [MBF] is objective and can be used in lieu of the *P* values as a measure of the evidential strength." The MBF does *not* require specification of prior probabilities. It is calculated, in a comparison of two group means, as the ratio of the likelihood function at two points, $\partial = 0$ and $\partial = d$, where *d* is the observed difference between the sample means. Goodman shows that if the data are such that a significance test yields $P = 0.05$, then the MBF will equal 0.15, "meaning that the null hypothesis gets 15% as much support as the single best supported [point alternative] hypothesis … [thus] indicating that the evidence against the null hypothesis is not nearly as strong as '$P = 0.05$' suggests."

A better logician will counter "So what?" to the first part of that statement and "*non sequitur* and false!" to the second. Let's imagine that $d = 12.2$. In almost any real situation in the basic or applied sciences, the investigator will want to report the actual means, *d*, *P*, samples sizes and possibly confidence intervals or other auxiliary descriptors of the data set. But a measure of the likelihood of the null relative to the likelihood of the point alternative that $\partial = 12.2$ will be of no value or interest. The investigator and his readers have no focused interest in the likelihood of that specific ∂ of 12.2, but only in having some confidence that $\partial \neq 0$ and that the sign and magnitude of and precision of *d* are reasonably estimated by some standard. Goodman (1999b) claims that "many Bayesian reanalyses of clinical trials conclude that the observed differences are not likely to be true" despite, he implies, being supported by low (< 0.05) *P* values. His statement is true, of course, if it is taken to mean only that *d* is rarely likely to equal ∂. In a similar vein, Lindley (1990) raises the question of "how much money has been wasted on inappropriate, incoherent analyses of clinical trials" because Bayesian methods were not used. We can only exclaim, "Woe to medicine in the clutches of Bayesians!" In a gently worded but devastating critique, Moyé (2008) concurs. He summarizes the many philosophical and practical problems that prevent Bayesian methods, in their current state, from having much positive value for the analysis of clinical trials or, by extension, manipulative experiments of any sort.

Pluralism remembered

Clarification of this confusion requires only an understanding of the differing purposes and capabilities of different methodologies. A *P* value is a measure of the *absolute* plausibility of H_0. We generally like this to be low so we can get on to analysis of what the estimated effect size means for the phenomena under study. The null and alternative hypotheses are not scientific or research hypotheses, and their relative plausibilities matter but are not a main concern. Focus

is on estimation. In most significance assessment situations, effect size is indeed estimated as one step in the calculation of the *P* value. (Above, the qualifier *absolute* is used to emphasise that *P* is not the plausibility of H_0 relative to any particular point or subset of alternative hypotheses within H_1: $\partial \neq 0$. The absolute plausibility of H_0 can be higher or lower than the absolute plausibility that might be calculated for any point H_1 by setting that H_1 up as the null. A *P* value is *not* a measure of the relative or absolute implausibility of the composite H_1: $\partial \neq 0$, as that H_1 is plausible whether *P* is high or low. And *P* certainly is *not* the probability that H_0 is true).

On the other side of a tall fence are other methodologies appropriate to those less common situations where interest is in assessing the *relative* plausibilities of two or more point hypotheses or models. Bayesian statistics, as well as likelihood and information theoretic methodologies, are available for these objectives. Some persons on this multi-model side of the fence often seem not to notice the fence. They are given to making statements such as: "In view of the likelihood principle, all of these classical [frequentist] ideas come under new scrutiny, and must, I believe, be abandoned or seriously modified" (Savage 1962: 18); "… the reason why a plausible rationale for signifcance tests has not yet been found is because none exists" (Royall 1997: 68); "The solution to the problem of statistical inference … is to switch from the *p* value methodology to a model selection methodology" (Wagenmakers 2007); or "I recommend that ecologists largely stop using [significance tests] in favor of [Bayesian and information-theoretic methods] (McCarthy 2007)."

Fortunately, there are appearing more and more papers that accept the obvious fact that significance assessment and other statistical methodologies serve different functions and can coexist in the arsenal of any individual scientist or statistician. Among works evincing this attitude in various ways we can mention Levin (1998b), McLean and Ernest (1998), Howard *et al*. (2000), Gigerenzer *et al*. (2004), Clark (2005), Stephens *et al*. (2005), Cox (2006a), and Hobbs and Hilborn (2006). Surely many calm souls regard the fact as so obvious to be hardly worth stating.

P values based on "unobserved data"?

Basing conclusions on 'unobserved data' would, on the face of it, not sound like a good idea in any circumstance. But this has been a pejorative manner of describing *P* values that has been much used by *aficionados* of Bayesian and likelihood methods in their critiques of frequentist methods. In his classic early treatise on likelihood methods, Jeffreys (1939: 319) stated that "The use of the *P* integral in significance tests … is fallacious because it rejects the hypothesis on account of observations that have not occurred." This has been repeated over and over in misleading and uncritical fashion. *P* values are said to be based on: "additional, unlikely and unobserved results" (Ellison 1996); "unobserved values" (Royall 1997); "data that were not observed" (Johnson 1999); "less likely, unobserved results" (Anderson *et al*. 2000); or "data that were never observed" (Wagenmakers 2007).

Such remarks refer to the fact that *P* is not the probability of the observed 'point' result, e.g. d, given H_0, but rather the probability of that result or a more 'extreme' one, e.g. a result $\geq |d|$ in a standard two-tailed *t*-test. It is the probability of a class of hypothetical results *but a class that is completely defined by the observed data or result*. There are no "unobserved data" at issue. Defined in this classic Fisherian manner, most scientists and statisticians reasonably regard a low *P* value as good evidence against H_0 and in favor of H_1 on the simple ground that the only other way to account for a low *P* value would be to suppose that a rare (or very rare, or very, very rare) event had occurred, i.e. one in the class of all values $\geq |d|$ when $\partial = 0$. Bayesians and likelihooders sometimes explain their objection to *P* values by saying they violate the likelihood principle. But this only means that those folks object to the question the neoFisherian frequentist is asking and demand that he ask another, even if that other question is not of interest or cannot be answered well.

Final remarks

The phrase "final collapse" in our title acknowledges the work of armies of critics who have

gone before us. Their cannonades had done most of the job of taking down the baroque cathedral of paleoFisherian-Neyman-Pearsonian statistical catechisms, and many of its inhabitants had fled long ago to a simpler but better-ordered and Bayesian-proof, neoFisherian cottage down the road. We came along after the dust had settled, and have just tried to push over the last remaining structures of the old cathedral and to show the logic of the neoFisherian reformation. Most of the stone building blocks from the old cathedral were still of value. They just needed to be reassembled with fresh mortar by a new generation of scientists and statisticians to increase the guest capacity and beautify the gardens of the neoFisherian cottage. At the end of the lane, some new neighbors have their own cottage nearly finished, constructed with modern likelihood and information theoretic tools. The more temperate of these folks should be a nice addition to local society, especially if their dogs are kept out of the neoFisherian petunia beds.

Acknowledgments

This work was supported in part by a grant from the Association for the Study of Animal Behaviour to CML. CML also thanks San Diego State University for its hospitality and use of its facilities during her study visit there. For helpful comments of various versions of this article, we express great thanks to Nekane Balluerka, Siu L. Chow, David R. Cox, Brian Dennis, Fiona M. Fidler, Emili García-Berthou, Steven Goodman, Bob O'Hara, Richard J. Harris, Allen H. Hurlbert, Deborah G. Mayo, Michael Riggs, Scott Roesch, Keith Sockman, Aris Spanos, N. Scott Urquhart, and a number of spirited anonymous reviewers. Some ideas discussed in this paper were presented in a talk by SH at the annual meeting of The Finnish Society of Forest Science in October 2002. Hannu Rita is thanked for arranging that invitation and the Finnish foresters for the warm reception they accorded the first announcement of the formation of the *NeoFisherian Statistical Liberation Front* made at the end of that talk.

References

Abelson, R. P. 1995: *Statistics as principled argument*. — Lawrence Erlbaum, Hillsdale, New Jersey.

Abelson, R. P. 1997: A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). — In: Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (eds.), *What if there were no significance tests?*: 117–144. Laurence Erlbaum Associates, Mahwah, New Jersey.

Altman, D. G. 1991: *Practical statistics for medical research*. — Chapman and Hall, London.

Altman, D. G., Gore, S. M., Gardner, M. J. & Pocock, S. J. 1983: Statistical guidelines for medical journals. — *British Medical Journal* 286: 1489–1493.

Altman, D. G., Machin, D., Bryant, T. N. & Gardner, M. J. 2000: *Statistics with confidence*, 2nd ed. — British Medical Journal, Arrowsmith, Bristol UK.

Anderson, D. R., Burnham, K. P. & Thompson, W. L. 2000: Null hypothesis testing: problems, prevalence, and an alternative. — *Journal of Wildlife Management* 64: 912–923.

Anderson, T. W. 1987: Comment on "A review of multivariate analysis" by M. J. Schervish. — *Statistical Science* 2: 413–417.

Animal Behaviour 2009: Guide for authors [submitting manuscripts to *Animal Behaviour*]. — http://www.elsevier.com/wps/find/journaldescription.cws_home/622782/authorinstructions.

Anscombe, F. J. 1961: Examination of residuals. — *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1: 1–35.

APA 2001: *Publication manual of the American Psychological Association*, 5th ed. — APA, Washington, DC.

Bahadur, R. R. 1952: A property of the *t*-statistic. — *Sankhya* 12: 79–88.

Balluerka, N., Gomez, J. & Hidalgo, D. 2005: The controversy over null hypothesis significance testing revisited. — *Methodology* 1: 55–70.

Barnard, G. A. 1982: Conditionality versus similarity in the analysis of 2 × 2 tables. — In: Rao, C. R., Kallianpur, G., Krishnaiah, P. R. & Ghosh, J. K. (eds.), *Statistics and probability: Essays in honor of C. R. Rao*: 59–65. North Holland Publ., New York.

Bart, J., Fligner, M. A. & Notz, W. I. 1998: *Sampling and statistical methods for behavioral ecologists*. — Cambridge University Press, Cambridge.

Bender, R., Berg, G. & Zeeb, H. 2005: Tutorial: using confidence intervals in medical research. — *Biometrical Journal* 47: 237–247.

Berger, J. O. 2003: Could Fisher, Jeffreys and Neyman have agreed on testing? — *Statistical Science* 18: 1–32.

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

Berger, J. O. & Delampady, M. 1987: Testing precise hypotheses. — *Statistical Science* 2: 317–352.

Berkson, J. 1938: Some difficulties encountered in the application of the Chi-square test. — *Journal of the American Statistical Association* 33: 526–542.

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

Birnbaum, A. 1977: The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley-Savage argument for Bayesian theory. — *Synthese* 36: 19–49.

Bolles, R. C. 1962: The difference between statistical hypoth-

eses and scientific hypotheses. — *Psychological Reports* 11: 639–645.

Boruch, R. 2007: The null hypothesis is not called that for nothing: statistical tests in randomized trials. — *Journal of Experimental Criminology* 3: 1–20.

Buchanan-Wollaston, H. J. 1935: Statistical tests. — *Nature* 136: 182–183.

Burke, C. J. 1953: A brief note on one-tailed tests. — *Psychological Bulletin* 50: 384–387.

Camilli, G. 1990: The test of homogeneity for 2 × 2 contingency tables: a review of and some personal opinions on the controversy. — *Psychological Bulletin* 108: 135–145.

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

Casella, G. & Berger, R. L. 1987: Comment on "Testing precise hypotheses" by J. O. Berger & M. Delampady. — *Statistical Science* 2: 344–347.

Caskey, L. L., Riedel, R. R., Costa-Pierce, B., Butler, J. & Hurlbert, S. H. 2007: Abundance, growth, and distribution of tilapia (*Oreochromis mossambicus*) in the Salton Sea, 1999–2002, with notes on bairdiella (*Bairdiella icistia)* and orangemouth corvina (*Cynoscion xanthulus)*. — *Hydrobiologia* 576: 185–203.

Chow, S. L. 1987: Science, ecological validity and experimentation. — *Journal for the Theory of Social Behavior* 17: 2–11.

Chow, S. L. 1996: *Statistical significance: rationale, validity and utility*. — Sage, Beverly Hills, California.

Christensen, R. 2005: Testing Fisher, Neyman, Pearson, and Bayes. — *American Statistician* 59: 121–126.

Clark, C. A. 1963: Hypothesis testing in relation to statistical methodology. — *Review of Educational Research* 33: 455–473.

Clark, J. S. 2005: Why environmental scientists are becoming Bayesians. — *Ecology Letters* 8: 2–14.

Cohen, J. 1990: Things I have learned (so far). — *American Psychologist* 45: 304–1312.

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

Copi, I. M. 1953: *Introduction to logic*. — J. H. Mcmillan, New York.

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

Cowles, M. 1989: *Statistics in psychology: an historical perspective*. — Lawrence Erlbaum, Hillsdale, New Jersey.

Cowles, M. & Davis, C. 1982: On the origins of the .05 level of statistical significance. — *American Psychologist* 37: 553–558.

Cox, D. R. 1958: Some problems connected with statistical inference. — *Annals of Mathematical Statistics* 29: 357–372.

Cox, D. R. 1977: The role of significance tests. — *Scandinavian Journal of Statistics* 4: 49–70.

Cox, D. R. 2006a: *Principles of statistical inference*. — Cambridge University Press, Cambridge.

Cox, D. R. 2006b: Frequentist and Bayesian statistics: a critique. — In: *Statistical problems in particle physics, astrophysics and cosmology*: 1–4. Imperial College Press, London.

Cumming, G. & Fidler, F. 2009: Confidence intervals: better answers to better questions. — *Journal of Psychology* 217: 15–26.

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

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., Lo, J., McMenamin, N. & Wilson, S. 2007: Statistical reform in psychology: is anything changing? — *Psychological Science* 18: 230–232.

Daniel, L. G. 1998: Statistical significance testing: a historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. — *Research in the Schools* 5: 23–32.

Daniel, W. W. 1990: *Applied nonparametric statistics*, 2nd ed. — PWS-KENT Publ. Co., Boston.

Dar, R., Serlin, R. C. & Omer, H. 1994: Misuse of statistical tests in three decades of psychotherapy research. — *Journal of Consulting and Clinical Psychology* 62: 75–82.

Dawkins, R. 2006: *The god delusion*. — Houghton Mifflin, Boston.

Dennis, B. 1996: Discussion: Should ecologists become Bayesians? — *Ecological Applications* 6: 1095–1103.

Dennis, B. 2004: Statistics and the scientific method in ecology. — In: Taper, M. L. & Lele, S. R. (eds.), *The nature of scientific evidence*: 327–378. University of Chicago Press, Chicago.

Detwiler, P. M., Coe, M. F. & Dexter, D. M. 2002: The benthic invertebrates of the Salton Sea: distribution and seasonal dynamics. — *Hydrobiologia* 473: 139–160.

Di Stefano, J. 2004: A confidence interval approach to data analysis. — *Forest Ecology and Management* 187: 173–183.

Eberhardt, L. 2003: What should we do about hypothesis testing? — *Journal of Wildlife Management* 70: 241–247.

Edwards, W., Lindeman, H. & Savage, L. J. 1963: Bayesian statistical inference for psychological research. — *Psychological Review* 70: 193–240.

Efron, B. 1986: Why isn't everyone a Bayesian? — *American Statistician* 40: 1–11.

Elenbaas, R. M., Elenbaas, J. A. & Cuddy, P. G. 1983: Evaluating the medical literature, Part II: statistical analysis. — *Annals of Emergency Medicine* 12: 610–620.

Ellison, A. M. 1996: An introduction to Bayesian inference for ecological research and environmental decision-making. — *Ecological Applications* 6: 1036–1046.

Ellison, A. M. 2004: Bayesian inference in ecology. — *Ecology Letters* 7: 509–520.

ESA 2006: *Guidelines for statistical analysis and data presentation*. — Ecological Society of America, available at http://esapubs.org/esapubs/preparation.htm.

Estes, W. K. 1997: Significance testing in psychological research: some persisting issues. — *Psychological Science* 8: 18–20.

Eysenck, H. J. 1960: The concept of statistical significance and the controversy about one-tailed tests. — *Psychological Review* 67: 269–271.

Falk, R. & Greenbaum, C. W. 1995: Significance tests die hard. — *Theory and Psychology* 5: 75–98.

Feinstein, A. R. 1975: Clinical biostatistics, XXVII: Biological dependency, 'hypothesis testing', unilateral probabilities, and other issues in scientific direction *vs*. statistical duplexity. — *Clinical Pharmacology and Therapeutics* 17: 499–513.

Fidler, F. M. 2002: The fifth edition of the APA Publication Manual: why its statistics recommendations are so controversial. — *Educational and Psychological Measurement* 62: 749–770.

Fidler, F. M. & Cumming, G. 2008: The new stats: attitudes for the twenty-first century. — In: Osborne, J. W. (ed.), *Best practice in quantitative methods*: 1–12. Sage, Thousand Oaks, California.

Fidler, F., Cumming, G., Burgman, M. & Thomason, N. 2004a: Statistical reform in medicine, psychology and ecology. — *Journal of Socio-Economics* 32: 615–630.

Fidler, F., Thomason, N., Cumming, G., Finch, S. & Leeman, J. 2004b: Editors can lead researchers to confidence intervals, but can't make them think. — *Psychological Science* 15: 119–126.

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

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., Edmonds, H., Harrington, C. & Schmitt, R. 2005b: Towards improved statistical reporting in the Journal of Consulting and Clinical Psychology. — *Journal of Consulting and Clinical Psychology* 73: 136–143.

Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R. & Thomason, N. 2006: Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. — *Conservation Biology* 20: 1539–1544.

Fisher, R. A. 1922: On the mathematical foundations of theoretical statistics. — *Philosophical Transactions of the Royal Society of London A* 222: 309–368.

Fisher, R. A. 1925, 1930, 1958: *Statistical methods for research workers* (1st, 3rd, 13th editions). — Oliver and Boyd, Edinburgh.

Fisher, R. A. 1935a, 1960: *The design of experiments* (1st ed., 1935; 7th ed., 1960). — Oliver and Boyd, Edinburgh.

Fisher, R. A. 1935b: Statistical tests. — *Nature* 136: 474.

Fisher, R. A. 1935c: The logic of inductive inference. — *Journal of the Royal Statistical Society* 98: 71–76.

Fisher, R. A. 1955: Statistical methods and scientific induction. — *Journal of the Royal Statistical Society B* 17: 69–78.

Fisher, R. A. 1956: *Statistical methods and scientific inference*. — Oliver and Boyd, London.

Fleiss, J. L. 1981: *Statistical methods for rates and proportions*, 2nd ed. — Wiley, New York.

Fleiss, J. L. 1986: *The design and analysis of clinical experiments*. — Wiley and Sons, New York.

Fleiss, J. L. 1987: Some thoughts on two-tailed tests. — *Journal of Controlled Clinical Trials* 8: 394.

Ford, E. D. 2000: *Scientific method for ecological research*. — Cambridge University Press, Cambridge.

Freedman, D., Pisani, R., Purves, R. & Adhikari, A. 1991: *Statistics*, 2nd ed. — Norton, New York.

Frick, R. W. 1995: Accepting the null hypothesis. — *Memory and Cognition* 23: 132–138.

Frick. R. W. 1996: The appropriate use of null hypothesis testing. — *Psychological Methods* 1: 379–390.

Gardner, M. J. & Altman, D. G. 1989: *Statistics with confidence*. — British Medical Journal, London.

Gibbons, J. D. & Pratt, J. W. 1975: *P*-values: interpretation and methodology. — *American Statistician* 29: 20–25.

Gigerenzer, G. 1993: The superego, the ego and the id in statistical reasoning. — In: Keren, G. & Lewis, C. (eds.), *A handbook for data aalysis in the behavioral sciences: methodological issues*: 311–339. Lawrence Erlbaum, Hillsdale, New Jersey.

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: Kaplan, D. (ed.), *The Sage handbook of quantitative methodology for the social sciences*: 391–408. Sage, Thousand Oaks, California.

Gigerenzer, G. & Murray, D. J. 1987: *Cognition as intuitive statistics*. — Lawrence Erlbaum, Hillsdale, New Jersey.

Goldfried, M. R. 1959: One-tailed tests and "unexpected" results. — *Psychological Review* 66: 79–80.

Good, I. J. 1982: Comments, conjectures and conclusions. — *Journal of Statistical Computation and Simulation* 16: 65–66.

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

Goodman, S. N. 1999b: Toward evidence-based medical statistics. 2: The Bayes factor. — *Annals of Internal Medicine* 130: 1005–1013.

Goodman, S. N. 2003: Commentary: The *P*-value, devalued. — *International Journal of Epidemiology* 32: 699–702.

Gotelli, N. J. & Ellison, A. M. 2004: *A primer of ecological statistics*. — Sinauer, Sunderland, Massachusetts.

Gotelli, N. J. & McGill, B. J. 2006: Null models versus neutral models: what's the difference? — *Ecography* 29: 793–800.

Graham, M. H. & Edwards, M. S. 2001: Statistical significance *vs*. fit: estimating the relative importance of individual factors in ecological analysis of variance. — *Oikos* 93: 503–513.

Grant, D. A. 1962: Testing the null hypothesis and the strategy and tactics of investigating theoretical models. — *Psychological Review* 69: 54–61.

Greenland, S. 1998: Meta-analysis. — In: Rothman, K. & Greenland, S. (eds.), *Modern epidemiology*, 2nd ed: 643–673. Lippincott-Raven, Philadelphia.

Greenwald, G. M. & Hurlbert, S. H. 1993: Microcosm analysis of salinity effects on coastal lagoon plankton assemblages. — *Hydrobiologia* 267: 307–335.

Guthery, F. S., Lusk, J. J. & Peterson, M. K. 2001: The fall of the null hypothesis: liabilities and opportunities. — *Journal of Wildlife Management* 65: 379–384.

Guttman, L. 1985: The illogic of statistical inference for cumulative science. — *Applied stochastic models and data analysis* 1: 3–10.

Hagen, R. L. 1997: In praise of the null hypothesis significance test. — *American Psychologist* 52: 15–24.

Hagen, R. L. 1998: A further look at wrong reasons to abandon statistical testing. — *American Psychologist* 53: 801–803.

Hagood, M. J. 1941: *Statistics for sociologists*. — Henry Holt & Co., New York.

Hall, P. & Selinger, B. 1986: Statistical significance: balancing evidence against doubt. — *Australian Journal of Statistics* 28: 354–370.

Harris, R. J. 1997a: Reforming significance testing via three-valued logic. — In: Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (eds.), *What if there were no significance tests?*: 145–174. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Harris, R. J. 1997b: Significance tests have their place. — *Psychological Science* 8: 8–11.

Harris, R. J. 2005: Classical statistical inference: practice versus presentation. — *Encyclopedia of statistics in behavioral science* 1: 268–278.

Hart, C. M., Gonzalez, M. R., Simpson, E. P. & Hurlbert, S. H. 1998: Salinity and fish effects on Salton Sea microecosystems: zooplankton and nekton. — *Hydrobiologia* 381: 129–152.

Hawkins, D. 2005: *Biomeasurement*. — Oxford Univ. Press, New York.

Henkel, R. E. 1976: *Tests of significance*. — Sage Publications, Beverly Hills, California.

Henny, C. J., Anderson, T. W. & Crayon, J. J. 2007: Organochlorine pesticides, polychlorinated biphenyls, metals, and trace elements in waterbird eggs, Salton Sea, California, 2004. — *Hydrobiologia* 604: 137–149.

Hill, T. & Lewicki, P. 2007: *Statistics: methods and applications*. — StatSoft, Tulsa, Oklahoma.

Hobbs, N. T. & Hilborn, R. 2006: Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. — *Ecological Applications* 16: 5–19.

Hogben, L. 1957: *Statistical theory: the relationship of probability, credibility and error*. — Norton, New York.

Hoover, D. K. & Siegler, M. V. 2008: Sound and fury: McCloskey and significance testing in economics. *Journal of Economic Methodology* 15: 1–37.

Howard, G. S., Maxwell, S. E. & Fleming, K. J. 2000: The proof of the pudding: an illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. — *Psychological Methods* 5: 315–332.

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

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.

Huberty, G. J. 1993: Historical origins of statistical testing practices: the treatment of Fisher versus Neyman-Pearson views in textbooks. — *Journal of Experimental Education* 61: 317–333.

Hunter, J. E. 1997: Needed: a ban on the significance test. — *Psychological Science* 8: 3–7.

Hurlbert, S. H. 1994: Old shibboleths and new syntheses [Review of *Design and analysis of ecological experiments*, Scheiner, S. M. & Gurevitch, J., eds.]. — *Trends in Ecology and Evolution* 9: 495–496.

Hurlbert, S. H. 1998: Experiments in ecology [Review of *Experiments in Ecology* by Underwood, A. J.]. — *Endeavour* 21: 172–173.

Hurlbert, S. H. & Lombardi, C. M. 2003: Design and analysis: uncertain intent, uncertain result [Review of *Experimental design and data analysis for biologists*, by Quinn, G. P. & Keough, M. J.]. — *Ecology* 84: 810–812.

ICMJE 2007: *Uniform requirements for manuscripts submitted to biomedical journals: writing and editing for biomedical publication*. — International Committee of Medical Journal Editors, available at http://www.icmje.org.

Inman, H. F. 1994: Karl Pearson and R. A. Fisher on statistical tests: a 1935 exchange from *Nature*. — *The American Statistician* 48: 2–11.

Jacard, J. & Guilamo-Ranos, V. 2002: Analysis of variance frameworks in clinical child and adolescent psychology: advanced issues and recommendations. — *Journal of Clinical Child Psychology* 31: 278–294.

Jeffreys, H. 1939: *Theory of probability*. — Clarendon Press, Oxford.

Johnson, D. H. 1999: The insignificance of statistical significance testing. — *Journal of Wildlife Management* 63: 763–772.

Kaiser, H. F. 1960: Directional statistical decisions. — *Psychological Review* 67: 160–167.

Kalbfleisch, J. G. & Sprott, D. A. 1976: On 'tests of significance.' — In: Harper, W. L. & Hooker, C. A. (eds.), *Foundations of probability theory, statistical inference, and statistical theories of science*, vol. 2: 259–272. Reidel, Boston.

Kendall, M. G. 1963: Ronald Aylmer Fisher, 1890–1962. — *Biometrika* 50: 1–15.

Kendall, P. C. 1997: Editorial. — *Journal of Consulting and Clinical Psychology* 15: 3–5.

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

Kimmel, H. D. 1957: Three criteria for the use of one-tailed tests. — *Psychological Bulletin* 54: 351–353.

Kirk, R. E. 2007: Effect magnitude: a different focus. — *Journal of Statistical Planning and Inference* 137: 1634–1646.

Kline, R. B. 2004: *Beyond significance testing*. — American Psychological Association, Washington DC.

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

Lang, T. A. & Secic, M. 1997: *How to report statistics in medicine*. — American College of Physicians, Philadelphia.

Lecam, L. 1977: A note on metastatistics, or 'an essay toward stating a problem in the doctrine of chances'. — *Synthese* 36: 133–160.

Lecoutre, B., Lecoutre, M.-P. & Poitevineau, J. 2001: Uses, abuses and misuses of significance tests in the scientific community: won't the Bayesian choice be unavoidable? — *International Statistical Review* 69: 399–417.

Lehmann, E. L. 1950: Some principles of the theory of testing hypotheses. — *Annals of Mathematical Statistics* 21: 1–26.

Lehmann, E. L. 1959: *Testing statistical hypotheses*. — Springer, New York.

Lehmann, E. L. 1987: [Review of *Error and the growth of experimental knowledge* by D.G. Mayo]. — *Journal of the American Statistical Association* 92: 789.

Lehmann, E. L. 1993: The Fisher-Neyman-Pearson theories of testing hypotheses: one theory or two? — *Journal of the American Statistical Association* 88: 1242–1249.

Lehmann, E. L. & Romano, J. P. 2005: *Testing statistical hypotheses,* 3d ed. — Springer, New York.

Lenhard, J. 2006: Models and statistical inference: the controversy between Fisher and Neyman-Pearson. — *British Journal of the Philosophy of Science* 57: 69–91.

Lenth, R. V. 2001: Some practical guidelines for effective sample size determinations. — *American Statistician* 55: 187–193.

Leventhal, L. & Huynh, C.-L. 1996: Directional decisions for two-tailed tests: power, error rates, and sample size. — *Psychological Methods* 1: 278–292.

Levin, J. R. 1998a: What if there were no more bickering about statistical significance tests? — *Research in the Schools* 5: 43–53.

Levin, J. R. 1998b: To test or not to test H_0? — *Educational and Psychological Measurement* 58: 313–333.

Levine, T. R., Weber, R., Hullett, C., Park, H. S. & Massi-Lindsey, L. L. 2008a: A critical assessment of null hypothesis significance testing in quantitative communication research. — *Human Communication Research* 34: 171–187.

Levine, T. R., Weber, R., Park, H. S. & Hullett, C. 2008b: A communication researcher's guide to null hypothesis significance testing and alternatives. — *Human Communication Research* 34: 188–209.

Lindley, D. V. 1990: The 1998 Wald Memorial Lectures: The present position in Bayesian statistics. — *Statistical Science* 5: 44–89.

Lombardi, C. M. & Hurlbert, S. H. 2009: Misprescription and misuse of one-tailed tests. — *Austral Ecology* 34: 447–468.

Losavich, J. L., Neyman, J., Scott, E. L. & Wells, M. A. 1971: Hypothetical explanations of the negative apparent effects of cloud seeding in the Whitetop Experiment. — *Proceedings of the U.S. National Academy of Sciences* 68: 2643–2646.

Lukacs, P. M., Thompson, W. L., Kendall, W. L., Gould, W. R., Doherty, P. F., Burnham, K. P. & Anderson, D. R. 2007: Concerns regarding a call for pluralism of information theory and hypothesis testing. — *Journal of Applied Ecology* 44: 456–460.

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

Marks, R. G. 1982: *Designing a research project: the basics of biomedical research methodology*. — Lifetime Learning Publications, Belmont, California.

Mather, K. 1951: *Statistical analysis in biology*, 4th ed. — Methuen, London.

Mayo, D. G. 1985: Behavoiristic, evidentialist and learning models of statistical testing. — *Philosophy of Science* 52: 493–516.

Mayo, D. G. 1992: Did Pearson reject the Neyman-Pearson philosophy of statistics? — *Synthese* 90: 233–262.

Mayo, D. G. 1996: *Error and the growth of experimental knowledge*. — University of Chicago Press, Chicago.

Mayo, D. G. 1997: Error statistics and learning from error: making a virtue of necessity. — *Philosophy of Science* 64 (Proceedings): S195–S212.

Mayo, D. G. 2005: Philosophy of statistics. — In: Sarkar, S. & Pfeifer, J. (eds.), *Philosophy of science: an encyclopedia*: 802–815. Routledge, London.

Mayo, D. G. & Cox, D. R. 2006: Frequentist statistics as a theory of evidence. — In: *Optimality: The Second Erich L. Lehmann Symposium*: 77–97. Institute of Mathematical Statistics, Lecture Notes-Monograph Series, vol. 49, Ohio.

Mayo, D. G. & Spanos, A. 2006: Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. — *British Journal of the the Philosophy of Science* 57: 323–357.

Mayo, D. G. & Spanos, A. 2009: Error statistics. — In: Gabbay, D., Thagard, P. & Woods, J. (eds.), *Philosophy of statistics*. *Handbook of philosophy of science*. Elsevier, Amsterdam. [In press].

McCarthy, M. A. 2007: *Bayesian methods for ecology*. — Cambridge University Press, Cambridge.

McCulloch, C. M. 2004: Commentary [on *Statistics and the scientific method*, by B. Dennis]. — In: Taper, M. L. & Lele, S. R. (eds.), *The nature of scientific evidence*: 360–362. University of Chicago Press, Chicago.

McGinnis, R. 1958: Randomization and inference in sociological research. — *American Sociological Review* 23: 408–414.

McLean, J. E. & Ernest, J. E. 1998: The role of statistical significance testing in educational research. — *Research in the Schools* 5: 15–22.

Mead, R. & Curnow, R. N. 1983: *Statistical methods in agriculture and experimental biology*. — Chapman and Hall, New York.

Meehl, P. E. 1967: Theory-testing in psychology and physics: A methodological paradox. — *Philosophy of Science* 34: 103–115.

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

Melton, A. 1962: Editorial. — *Journal of Experimental Psychology* 64: 553–557.

Mills, S. R. 2003: Statistical practices: the seven deadly sins. — *Child Neuropsychology* 2003: 221–233.

Moore, D. S. 1985: *Statistics: concepts and controversies*, 2nd ed. — Freeman, New York.

Moreau, M. F., Surico-Bennett, J., Vicario-Fisher, M., Gerards, R., Gersberg, R. M. & Hurlbert, S. H. 2007: Selenium, arsenic, DDT, and other contaminants in four fish species in the Salton Sea, California, their temporal

trends and their potential impact on human consumers and wildlife. — *Lake and Reservoir Management* 23: 536–569.

Morrison, D. E. & Henkel, R. E. 1970: Significance tests in behavioral research: skeptical conclusions and beyond. — In: Morrison, D. E. & Henkel, R. E. (eds.), *The significance test controversy*: 305–311. Aldine, Chicago.

Moyé, L. A. 2008: Bayesians in clinical trials: asleep at the switch. — *Statistics in Medicine* 27: 469–482.

Mulaik, S. A., Raju, N. S. & Harshman, R. A. 1997: There is a time and place for significance testing. — In: Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (eds.), *What if there were no significance tests?*: 65–116. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Nakagawa, S. & Cuthill, I. C. 2007: Effect size, confidence interval and statistical significance: a practical guide for biologists. — *Biological Reviews* 82: 591–605.

Neyman, J. 1937: Outline of a theory of statistical estimation based on the classical theory of probability. — *Philosophical Transactions of the Royal Society A* 236: 333–380.

Neyman, J. 1950: *First course in probability and statistics*. — Henry Holt, New York.

Neyman, J. 1956: Note on an article by Sir Ronald Fisher. — *Journal of the Royal Statistical Society B* 18: 288–294.

Neyman, J. 1976: The emergence of mathematical statistics. — In: Owen, D. B. (ed.), *On the history of statistics and probability*: 149–193. Dekker, New York.

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

Neyman, J. & Pearson, E. S. 1928: On the use and interpretation of certain test criteria of statistical inference, part I. — *Biometrika* 20A: 175–240.

Neyman, J. & Pearson, E. S. 1933a: On the problem of the most efficient tests of statistical hypotheses. — *Philosophical Transactions of the Royal Society of London A* 231: 289–337.

Neyman, J. & Pearson, E. S. 1933b: The testing of statistical hypotheses in relation to probabilities a priori. — *Proceedings of the Cambridge Philosophical Society* 29: 492–510.

Nicholls, N. 2001: The insignificance of significance testing. — *Bulletin of the American Meteorological Society* 81: 981–986.

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

Nix, T. W. & Barnette, J. J. 1998: The data analysis dilemma: ban or a burden. A review of null hypothesis significance testing. — *Research in the Schools* 5: 3–14.

Nunnally, J. 1960: The place of statistics in psychology. — *Educational and Psychological Measurement* 20: 641–650.

Oakes, M. 1986: *Statistical inference: a commentary for the social and behavioural sciences*. — Wiley, New York.

Osenberg, C. W., Sarnelle, O. & Cooper, S. D. 1997: Effect size in ecological experiments: the application of biological models in meta-analysis. — *American Naturalist* 150: 798–812.

Osenberg, C. W., Sarnelle, O., Cooper, S. C. & Holt, R. D. 1999: Resolving ecological questions through meta-analysis: goals, metrics, and models. — *Ecology* 80: 1105–1117.

Pace, L. & Salvan, A. 1997: *Principles of statistical inference from a neo-Fisherian perspective*. — Advanced Series on Statistical Science and Applied Probability, vol. 4, World Scientific, Singapore.

Peace, K. E. 1989: The alternative hypothesis: one-sided or two-sided? — *Journal of Clinical Epidemiology* 42: 473–476.

Peace, K. E. 1991: One-sided or two-sided *P* values: which most appropriately address the question of drug efficacy? — *Journal of Biopharmaceutical Statistics* 1: 133–138.

Pearson, E. S. 1955: Statistical concepts in relation to reality. — *Journal of the Royal Statistical Society B* 17: 204–207.

Pearson, K. 1894: Contributions to the mathematical theory of evolution. — *Philosophical Transactions of the Royal Society A* 185: 71–110.

Pearson, K. (ed.) 1914: *Tables for statisticians and biometricians*. — The University Press, Cambridge.

Pearson, K. 1935a: Statistical tests. — *Nature* 136: 296–297.

Pearson, K. 1935b: Statistical tests. — *Nature* 136: 550.

Petraitis, P. 1998: How can we compare the importance of ecological processes if we never ask, "Compared to what?". — In: Resetarits, W. J. Jr. & Bernardo, J. (eds.), *Experimental ecology: issues and perspectives*: 183–201. Oxford University Press, Oxford.

Pillemer, D. B. 1991: One- *versus* two-tailed hypothesis tests in contemporary educational research. — *Educational Researcher* 20: 13–17.

Pollard, P. & Richardson, J. E. 1987: On the probability of making type I errors. — *Psychological Bulletin* 102: 159–163.

Poole, C. 1987: Beyond the confidence interval. — *American Journal of Public Health* 77: 195–199.

Quinn, G. P. & Keough, M. J. 2002: *Experimental design and data analysis for biologists*. — Cambridge University Press, New York.

Quinn, J. F. & Dunham, A. E. 1983: On hypothesis testing in ecology and evolution. — *American Naturalist* 122: 602–617.

Reichardt, C. S. & Golub, H. F. 1997: When confidence intervals should be used instead of statistical significance tests, and vice versa. — In: Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (eds.), *What if there were no significance tests?*: 259–286. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Reid, C. 1982: *Neyman — from life*. — Springer, New York.

Reifel, K. M., Trees, C. C., Olivo, E., Swan, B. K., Watts, J. M. & Hurlbert, S. H. 2007: Influence of river inflows on spatial variation of phytoplankton around the southern end of the Salton Sea, California. — *Hydrobiologia* 576: 167–183.

Rindskoff, D. M. 1997: Testing "small", not null hypotheses: classical and Bayesian approaches. — In: Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (eds.), *What if there were no significance tests?*: 319–332. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Rosenthal, R. & Rubin, D. B. 1994: The counternull value of an effect size: a new statistic. — *Psychological Science* 5: 329–334.

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

Rossi, J. S. 1997: A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. — In: Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (eds.), *What if there were no significance tests?*: 175–198. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Royall, R. M. 1997: *Statistical evidence: a likelihood paradigm*. — Chapman and Hall, New York.

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: Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (eds.), *What if there were no significance tests?*: 335–392. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Rutledge, T. & Loh, C. 2004: Effect sizes and statistical testing in the determination of clinical significance in behavioral research medicine. — *Annals of Behavioral Medicine* 27: 138–145.

Salsburg, D. S. 1985: The religion of statistics as practiced in medical journals. — *The American Statistician* 39: 220–223.

Salsburg, D. S. 1989: Use of restricted significance tests in clinical trials: Beyond the one- *versus* two-tailed controversy. — *Controlled Clinical Trials* 10: 71–82.

Salsburg, D. S. 1992: *The use of restricted significance tests in clinical trials*. — Springer, New York.

Salsburg, D. S. 1993: The use of statistical methods in the analysis of clinical studies. — *Journal of Clinical Epidemiology* 46: 17–27.

Sardella, B. A., Matey, V. & Brauner, C. J. 2007: Coping with multiple stressors: physiological mechanisms and strategies in fishes of the Salton Sea. — *Lake and Reservoir Management* 23: 518–527.

Savage, L. J. 1957: Nonparametric statistics [review of *Nonparametric statistics for the behavioral sciences*, by Siegel, S.]. — *Journal of the American Statistical Association* 52: 331–344.

Savage, L. J. (ed.) 1962: *The foundations of statistical inference: a discussion*. — Methuen, London.

Savage, L. J. 1976: On rereading R. A. Fisher. — *Annals of Statistics* 4: 441–500.

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

Schmidt, F. L. & Hunter, J. E. 1997: Eight common but false objections to the discontinuation of significance testing in the analysis of research data. — In: Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (eds.), *What if there were no significance tests?*: 37–64. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Schulman, R. S. 1992: *Statistics in plain English*. — Van Nostrand Reinhold, New York.

Schweder, T. 2003: Abundance estimation from multiple photo surveys: confidence distributions and reduced likelihoods for bowhead whales off Alaska. — *Biometrics* 59: 974–983.

Schweder, T. & Hjort, N. L. 2002: Confidence and likelihood. — *Scandinavian Journal of Statistics* 29: 309–322.

Sedlmeier, P. & Gigerenzer, G. 1989: Do studies of statistical power have an effect on the power of studies? — *Psychological Bulletin* 105: 309–316.

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

Serlin, R. C. & Lapsley, D. K. 1985: The good-enough principle. — *American Psychologist* 40: 73–83.

Siegel, S. 1956: *Nonparametric statistics for the behavioral sciences*. — McGraw-Hill, New York.

Shaver, J. P. 1993: What statistical significance testing is, and what it is not. — *Journal of Experimental Education* 61: 293–316.

Simberloff, D. 1990: Hypotheses, errors, and statistical assumptions. — *Herpetologica* 46: 351–357.

Simon, R. 1986: Confidence intervals for reporting results of clinical trials. — *Annals of Internal Medicine* 105: 429–435.

Skipper, K. S. Jr., Guenther, A. L. & Nass, G. 1967: The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. — *American Sociologist* 2: 16–18.

Smith, C. A. B. 1962: Prepared contribution to L. J. Savage *et al*., *The Foundations of statistical inference*: 58–61. — Methuen, London.

Sockman, K. 2008: Ovulation order mediates a trade-off between pre-hatching and post-hatching viability in an altricial bird. — *Plos ONE* 3:e1785.

Sokal, R. R. & Rohlf, F. J. 1995: *Biometry*, 3rd ed. — Freeman, San Francisco.

Spanos, A. 1999: *Probability theory and statistical inference: econometric modeling with observational data*. — Cambridge University Press, Cambridge.

Spanos, A. 2008: [Review of S. T. Ziliak & D. N. McCloskey, *The cult of statistical significance: how the standard error costs us jobs, justice, and lives*]. — *Erasmus Journal for Philosophy and Economics* 1: 154–164.

StatSoft Inc. 2007: *Electronic statistics textbook*. — http://www.statsoft.com/textbook/stathome.html.

Steel, R. G. D., Torrie, J. H. & Dickey, D. A. 1997: *Principles and procedures of statistics*, 3rd ed. — McGraw-Hill, New York.

Stephens, P. A., Buskirk, S. W., Hayward, G. D. & Martinez del Rio, C. 2005: Information theory and hypothesis testing: a call for pluralism. — *Journal of Applied Ecology* 42: 4–12.

Stigler, S. 2000: The problematic unity of biometrics. — *Biometrics* 56: 272–277.

Stigler, S. 2005: Discussion of D. Denis. — *Journal de la Societé Française de Statistique* 145: 63–64.

Stigler, S. 2008: Fisher and the 5% level. — *Chance* 21: 12.

Stoehr, A. M. 1999: Are significance threshholds appropriate for the study of animal behaviour? — *Animal Behaviour* 57: F22–F25.

Strong, D. R. Jr. 1980: Null hypotheses in ecology. — *Synthese* 43: 271–285.

Strong, D. R. Jr., Simberloff, D., Abele, L. G. & Thistle, A. B. 1984: *Ecological communities: conceptual issues and the evidence*. — Princeton University Press, Princeton, New Jersey.

Student 1906–1907: On the error of counting with a hemacytometer. — *Biometrika* 5: 351–360.

Student 1908: The probable error of a mean. — *Biometrika* 6: 1–25.

Swan, B. K., Reifel, K. M., Tiffany, M. A., Watts, J. M. & Hurlbert, S. H. 2007: Spatial and temporal patterns of transparency and light attentuation in the Salton Sea, California, 1997–1999. — *Lake and Reservoir Management* 23: 653–662.

Taylor, B. L. & Gerrodette, T. 1993: The uses of statistical power in conservation biology: the vaquita and northern spotted owl. — *Conservation Biology* 7: 489–500.

Thompson, B. 1998: Statistical significance and effect size reporting: portrait of a possible future. — *Research in the Schools* 5: 33–38.

Thompson, B. 2006: Critique of *p*-values. — *International Statistical Review* 74: 1–14.

Timms, B. V. 1998: Further studies on saline lakes of the eastern Paroo, inland New South Wales, Australia. — *Hydrobiologia* 381: 31–42.

Tippett, L. H. C. 1931: *The methods of statistics*. — Williams and Norgate, London.

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. 1960: Conclusions *vs*. decisions. — *Technometrics* 2: 423–433.

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

Underwood, A. J. 1997: *Experiments in ecology*. — Blackwell, London.

Unwin, S. 2003: *The probability of god: a simple calculation that proves the ultimate truth*. — Crown and Forum, New York.

Vacha-Haase, T., Reetz, D. R., Thompson, B., Nilsson, J. E. & Lance, T. S. 2000: Reporting practices and APA editorial policies regarding statistical significance and effect size. — *Theory and Psychology* 10: 413–425.

Venn, J. 1866: *The logic of chance*. — Macmillan, London.

Venn, J. 1888: Cambridge anthropometry. — *Journal of Anthropological Institute* 18: 140–154.

Wagenmakers, E. J. 2007: A practical solution to the practical problem of *p* values. — *Psychonomic Bulletin and Review* 14: 779–804.

Wald, A. 1939: Contributions to the theory of statistical estimation and testing hypotheses. — *Annals of Mathematical Statistics* 10: 299–326.

Wald, A. 1950: *Statistical decision functions*. — Wiley, New York.

Walter, S. D. 1995: Methods of reporting statistical results from medical research studies. — *American Journal of Epidemiology* 141: 896–906.

Wang, C. 1993: *Sense and nonsense of statistical inference*. — Marcel Dekker, New York.

Ware, J. H., Mosteller, F. & Ingelfinger, J. A. 1986: *P* values. — In: Bailar, J. C. III & Mosteller, F. (eds.), *Medical uses of statistics*: 149–169. NEJM Books, Waltham, Massachusetts.

Welkowitz, J., Cohen, B. H. & Ewen, R. B. 2006: *Introductory statistics for the behavioral sciences*, 6th ed. — Wiley, New York.

Welkowitz, J., Ewen, R. B. & Cohen, J. 1971 & 1991: *Introductory statistics for the behavioral sciences*, 1st and 4th editions. — Harcourt Brace Jovanovich, New York.

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

Wolterbeek, R. 1994: Statistical hypothesis should be brought in line with clinical hypothesis. — *British Medical Journal* 309: 873–874.

Yancey, J. M. 1996: Ten rules for reading clinical research reports. — *American Journal of Orthodontics and Dentofacial Orthopedics* 109: 558–564.

Yates, R. 1951: The influence of *Statistical methods for research workers* on the development of the science of statistics. — *Journal of the American Statistical Association* 46: 19–34.

Yates, F. 1984: Tests of significance for 2 × 2 contingency tables. — *Journal of the Royal Statistical Society A* 147 (part 3): 426–463.

Zar, J. H. 2004: *Biostatistical analysis*, 5th ed. — Prentice-Hall, Inc., New York.

Ziliak, S. T. & McCloskey, D. N. 2008: *The cult of statistical significance: how the standard error cost us jobs, justice and lives*. — University of Michigan Press, Ann Arbor.