Australian & New Zealand Journal of Statistics

*Aust. N. Z. J. Stat.* **54**(1), 2012, 23–42 doi: 10.1111/j.1467-842X.2012.00652.x

# LOPSIDED REASONING ON LOPSIDED TESTS AND MULTIPLE COMPARISONS

STUART H. HURLBERT[1,*] AND CELIA M. LOMBARDI[2]

*San Diego State University*
*and Consejo Nacional de Investigaciones Científicas y Técnicas*

## Summary

For those who have not recognized the disparate natures of tests of statistical hypotheses and tests of scientific hypotheses, one-tailed statistical tests of null hypotheses such as $\partial \leq 0$ or $\partial \geq 0$ have often seemed a reasonable procedure. We earlier reviewed the many grounds for not regarding them as such. To have at least some power for detection of effects in the unpredicted direction, several authors have independently proposed the use of lopsided (also termed split-tailed, directed or one-and-a-half-tailed) tests, two-tailed tests with $\alpha$ partitioned unequally between the two tails of the test statistic distribution. We review the history of these proposals and conclude that lopsided tests are never justified. They are based on the same misunderstandings that have led to massive misuse of one-tailed tests as well as to much needless worry, for more than half a century, over the various so-called 'multiplicity problems'. We discuss from a neo-Fisherian point of view the undesirable properties of multiple comparison procedures based on either (i) maximum potential set-wise (or family-wise) type I error rates (SWERs), or (ii) the increasingly fashionable, maximum potential false discovery rates (FDRs). Neither the classical nor the newer multiple comparison procedures based on fixed maximum potential set-wise error rates are helpful to the cogent analysis and interpretation of scientific data.

*Key words:* Bonferroni procedure; comparison-wise error rate; directed tests; false discovery rate; family-wise error rate; multiplicity; one-tailed tests; randomized clinical trials; set-wise error rate; significance tests; split-tailed tests; type I error.

## 1. Introduction

This review touches lightly but provocatively on several interconnected statistical controversies. We start by briefly summarizing our recent review (Lombardi & Hurlbert 2009) of the historically disparate advice as to when one-tailed tests are appropriate and the strong logical arguments against their use. We then describe several independent proposals for the use of 'lopsided' tests, namely two-tailed tests where $\alpha$, the maximum probability of a type I error, is allocated unequally to the two tails of the test statistic distribution. Such a procedure is thought by some to combine the best features of standard one- and two-tailed tests. Lopsided tests are now being promoted as one component of complex alpha-allocation or

*Author to whom correspondence should be addressed.
[1]Department of Biology, San Diego State University, San Diego, CA 92182-4614, USA.
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
*Acknowledgments.* For their suggestions on drafts of this article we thank Nekane Balluerka, Jarrett Byrnes, Roger Mead, Paul Murtaugh, Daniel O'Keefe, Michael Riggs, David Savitz, Allan Stewart-Oaten, Sheela Talwalker, Scott Urquhart, Alan Welsh – and three anonymous reviewers.

alpha-spending schemes for dealing with the 'multiplicity problem' in randomized clinical trials. This leads us to a brief critical review of those schemes, both the classical multiple comparison methods involving set-wise error rate procedures (SWERPs) and newer ones such as false discovery rate procedures (FDRPs), and of the general fear of high hypothetical set-wise error rates that drives this whole area of concern.

This critique is based on a neo-Fisherian paradigm or framework for significance assessment (Hurlbert & Lombardi 2009). Key elements of the framework are that, for individual tests: a comparison-wise type I error rate ($\alpha_C$ or CWER) is not specified, the terms 'significant' and 'non-significant' are not used, high *P*-values lead only to suspended judgment, more nuanced verbal characterization of results is demanded, and effect sizes and their precision are emphasized. In the following discussions we occasionally make reference to $\alpha_C$, but this is only for the purpose of critiquing other authors in the language of their own paleo-Fisherian or Neyman–Pearsonian frameworks. In the opinion of many, judicious interpretation of data and analyses is hindered by specification of $\alpha_C$, as discussed by Hurlbert & Lombardi (2009). The paleo-Fisherian and Neyman–Pearsonian frameworks both demand specification of $\alpha_C$. These differ primarily in that the paleo-Fisherian, like the neo-Fisherian, interprets high *P*-values as grounds for suspending judgment, whereas the Neyman–Pearsonian interprets high *P*-values as grounds for preferring the null over the alternative hypothesis.

## 2. One-tailed tests

Virtually all basic statistics textbooks introduce the concept and mechanics of one-tailed statistical procedures in addition to the more common and conventional two-tailed ones. One formerly popular reference work (Siegel 1956) advocated almost exclusive use of one-tailed procedures, a few simply encourage their use (e.g. Zar 2004), several recommend that one-tailed tests essentially never be used (Welkowitz *et al.* 1971, 1991; Fleiss 1986; Altman 1991; Schulman 1992; Bart *et al.* 1998; Hawkins 2005), and the majority give poor, vague or conflicting advice on the matter (e.g. Kendall & Stuart 1979; Kirk 1982; Glass & Hopkins 1984; Siegel & Castellan 1988; Moore & McCabe 1989; Darlington & Carson 1987; Snedecor & Cochran 1989; Sokal & Rohlf 1995; Underwood 1997; Kline 2004; Zar 2004; Martin & Bateson 2007). Lombardi & Hurlbert (2009) reviewed 52 statistics texts and concluded that only 12 of them – advising general avoidance of one-tailed tests – give reasonably good and clear advice.

Debate over when one-tailed tests are permissible has been going on for half a century. The logical arguments against their use, except in rare circumstances, in both basic and applied research (Kimmel 1957; Goldfried 1959; Welkowitz *et al.* 1971, 1991; Fleiss 1981, 1986; Oakes 1986; Cowles 1989; Altman 1991; Pillemer 1991) seem gradually to be winning out, at least in some disciplines. Lombardi & Hurlbert (2009) summarize and add to these arguments, and document that for two biological journals, *Animal Behaviour* and *Oecologia*, use of one-tailed tests apparently decreased by about 50 per cent between 1989 and 2005. The one category of 'rare circumstance' where one-tailed tests can be useful consists of those situations in which the null hypothesis states that a difference or effect size is greater or less than some particular non-zero value. Lombardi & Hurlbert (2009) briefly review such situations. Our further remarks here consider only the most common type of one-tailed test, where the null hypothesis is that $\partial \leq 0$ or $\partial \geq 0$, where $\partial$ represents the true difference between two means or any other true effect size.

The basic temptation that one-tailed tests offer is that they have slightly more power than two-tailed tests when the direction of the predicted effect is the same as that of the true effect. Their fundamental flaw is that they have zero power to detect effects in the direction opposite to that predicted. Thus, when results are obtained strongly in that opposite direction the investigator has only four options (Goldfried 1959; Lombardi & Hurlbert 2009). Option 1 is to report the value of $P$ (which will be $> 0.5$), argue that the unexpected result contains no information that is interesting to the investigator or science generally, and refrain from further analysis. Option 2 is to ignore the first value of $P$ and to calculate and report the $P$-value obtained in carrying out the two-tailed test on $H_0$: $\delta = 0$. Option 3 is to ignore the first value of $P$ and to calculate and report the $P$-value obtained in carrying out the one-tailed test for the unexpected direction. Finally, Option 4 is to acknowledge that there may be a real and interesting difference in the unexpected direction but to refrain from any testing for significance in that direction. As discussed by Lombardi & Hurlbert (2009), all four options – and hence all one-tailed tests involving nil nulls – are unacceptable. Options 1 and 2 represent a waste of resources. Options 2 and 3 represent *de facto* two-tailed tests that are likely to be reported deceptively, without mention of how 'horses were changed mid-stream'.

## 3. Unequal allocation: the best of both worlds?

Psychologists, starting in the early 1950s, were the first scientists to engage in fierce debate over the propriety of one-tailed tests. The earliest reasonably cogent proscription against their use was put forward by Kimmel (1957). But many are still confused, as evidenced, for example, by the vacillating treatment of one-tailed tests in the last three editions of the widely used psychology statistics text by Welkowitz *et al.* (1991, 1999, 2006; see discussion in Lombardi & Hurlbert 2009). Given this early ferment, it is not surprising that a psychologist also was apparently the first to suggest that the 'best of both worlds' might be obtained with a hybrid procedure that avoided having zero power for the detection of an unpredicted result. Kaiser (1960) raised the possibility, in any kind of two-tailed test, of apportioning $\alpha$ unequally between the two tails of the sampling distribution of the test statistic. Such tests would later be labelled 'split-tailed' (Braver 1975; Harris & Quade 1992; Harris 2005a, b), 'one-and-a-half-tailed' (Mantel 1983; Ramsey 1990), 'directed' (Rice & Gaines 1994c), or 'lopsided' (Abelson 1995). The last of these labels, with its clear connotation of asymmetry, seems most appropriate.

Shaffer (1972) thought that such tests 'might be desirable if it were more important to detect differences in one direction than in the other' but also recognized that 'it is rarely possible to quantify these considerations'. Braver (1975), however, argued that this 'unequal tails' approach might be a good compromise. He criticized those who advocated one-tailed tests for situations where the investigator expected the difference to be in a particular direction, noting that 'the insights to be gained from clear-cut results which are counter-intuitive or counter-theoretical sometimes exceed in importance those from findings that are consistent with predictions.' His recommendation was to assign 'the greater fraction [of $\alpha$] to the expected tail', acknowledging that 'a basic objection might be that such a procedure formalizes and thus sanctions the custom of giving less credence to disconfirming than confirming evidence, a practice which could be viewed as scientifically corrupt'. Braver then dismissed this objection by saying that this is 'the way all-too-human investigators presently act anyhow'.

We would say that such a practice would not be 'corrupt' so much as it would be arbitrary or, whenever unexpected results would be more important in some way than predicted ones, even unproductive. Thus Braver's arguments only lead us back to Shaffer's (1972) conclusion.

## 4. Multiple reinventions of the wheel

The idea of using unequally weighted tails has been independently re-proposed at least four or five times. Nosanchuk (1978) noted the 'simplistic and ambiguous' advice in most textbooks on when to use one-tailed tests and offered a tripartite procedure. If a 'process is of interest only if it yields findings in the given direction [expected]', he recommends the one-tailed test. '[F]or hypothesis-oriented research . . . where the researcher has clear expectations' he recommends the standard two-tailed test. But 'where the researcher has no clear expectations', Nosanchuk recommends an 'asymmetric' two-tailed test where $\alpha$ would be arbitrarily apportioned in the ratio of 9:1 between the 'expected outcome' and 'just-in-case' tails of the appropriate distribution. To compensate for insufficiently detailed statistical tables (in 1978) he says a ratio of 10:1 would also be acceptable. This would allow, for example, setting overall $\alpha$ at 0.055 and using respective tail areas of 0.05 and 0.005.

Meek & Turner (1983, p. 250) briefly mentioned the possibility of allocating $\alpha$ to the two tails in the ratio of 4:1 if one bound was 'considered more critical'. Meek & Ozgur (2004) urged that any two-semester course in statistics should cover lopsided tests and proposed that an $\alpha$-value should be selected and partitioned between the tails according to (i) the actual economic cost of a type I error and (ii) the relative costs of a negative versus a positive deviation from the null. Admitting that such cost information is almost never available they nevertheless recommended using 'subjective considerations' to implement lopsided partitioning of $\alpha$. In contrast, Harnett & Soni (1991, p. 270) considered the possibility of lopsided confidence intervals, recognized the usual absence of relevant cost information, and implied that such confidence intervals are best avoided.

Mantel (1983) and Ramsey (1990) acknowledged the undesirability of having zero power to detect unpredicted results and proposed 'one-and-a-half tailed' tests of significance. In Mantel's version, $\alpha$ would be apportioned in the ratio of 2:1 between the predicted and unpredicted tails for complex reasons relating to new approaches he was presenting on ordered alternatives. For Ramsey, $\alpha$ should be apportioned in the ratio of 4:1 between the predicted and unpredicted tails for the admittedly 'arbitrary' reason that when $\alpha = 0.05$, the critical tail area for the unpredicted direction would be 0.01, a value for $\alpha$ that is in 'widespread use'.

Later, Rice and Gaines came to a similar recommendation. In three papers (Gaines & Rice 1990; Rice & Gaines 1994a, b) they noted approvingly the frequent use of one-tailed tests by biologists where there are *a priori* expectations. Then they urged their increased use, in connection with tests for ordered alternatives, where more than two groups or treatments are being compared. In a fourth paper (Rice & Gaines 1994c), however, they changed course and recommended that two-tailed tests with unequally weighted tails 'be used in virtually all applications where one-sided tests have been previously used.' They referred to these as 'directed tests' and suggested, as had Ramsey (1990), that we adopt an arbitrary convention that the tails be weighted 4:1 in favour of the predicted direction of a result. They acknowledged that this reflected a philosophical position that 'we [should] require stronger empirical evidence to reject $H_0$ when the parameter differs in the unanticipated direction', but they offered no defence of that position.

Harris & Quade (1992) discussed 'split-tailed tests' and suggested that $\alpha$ allocation should generally favour the predicted tail, without being more specific. Later, Harris (2005a) presented an example where overall $\alpha$ is set at 0.05 with 0.04 allocated to the predicted tail and 0.01 to the other tail. He noted that unequal allocation will always result in a wider confidence interval and concluded that, while split-tailed tests are definitely preferable to one-tailed tests, two-tailed tests with equal allocation are much to be preferred over either. Likelihood-based confidence intervals, however, are asymmetric with equal allocation (e.g. Neale & Miller 1996) and so could actually be narrower with unequal allocation of $\alpha$.

Abelson (1995, pp. 55, 58) presented the case against the use of one-tailed tests, which he characterized as one of five common 'devices available to the desperate researcher for arguing that the results look good, when a dispassionate observer would say they are marginal or worse.' As a compromise, he suggested a 'lop-sided test' where the investigator wished to have greater power detecting effects in the predicted direction. In this procedure a result could be judged 'significant' at $\alpha = 0.055$ if a two-tailed test yields $P \leq 0.10$ with the result in the predicted direction or yields $P \leq 0.01$ with the result in the 'wrong' direction.

As Shaffer (1972) indicated, scientists are not likely to develop a consensus that detection of a predicted result is necessarily more important than detection of a contrary one, let alone a consensus as to what relative weighting of tails might be universally appropriate or acceptable as a convention. Lombardi & Hurlbert (2009) observed that Goldfried's (1959) Option 2, a decision procedure involving a one-tailed test automatically followed (when a result was in the unpredicted direction) by a classical two-tailed test, was the exact equivalent of a two-tailed test with overall $\alpha$ apportioned between the two tails in a ratio of 2:1, as recommended by Mantel (1983). While such weighting is, in a sense, more objective than the 4:1 weighting proposed by Ramsey (1990) and Rice & Gaines (1994c) or the 10:1 weighting proposed by Nosanchuk (1978) and Abelson (1995), it is still too arbitrary to be recommended. We note that Option 2, carried out without acknowledgment of its inflated $\alpha$, is what Abelson (1995, p.58) also 'wryly' called a 'one-and-a-half-tailed test'. This is different, of course, from the procedure to which Mantel (1983) and Ramsey (1990) applied that label.

Finally, Kornilov (1993) proposed a complex, semi-quantitative approach to this problem in which both relative expectations of *and* relative degrees of interest in different possible results determine whether a test should be one- or two-tailed and, if one-tailed, the direction to be tested. His procedure requires subjective quantification or weighting of four factors relating to expectations and interest. Given the rarity of interest in simple lopsided tests involving only one such subjective step – the partitioning of $\alpha$ among two tails – it is not surprising that Kornilov's approach has attracted little interest.

## 5. Alpha fixations in biomedical research

The whole notion of lopsided tests has an air of unreality about it. First, paleo-Fisherian and Neyman–Pearsonian authors set $\alpha$ at some arbitrary low value (e.g. 0.05) in order to be rigorous in their evidentiary standard, that is, to make it *difficult* to conclude that an effect exists when it does not. Those favouring lopsided testing then further arbitrarily engineer the test so that, if their prediction of the sign of a result is borne out, their chance of obtaining a $P$-value lower than the arbitrary $\alpha$ is somewhat greater than it would be with a simple two-tailed test; that is, they make it somewhat *easier* to conclude that there is an effect with a particular sign. Such gymnastics reflect an excessive preoccupation with statistical

hypothesis-testing and the completely unnecessary fixing of $\alpha$-levels for the conduct and interpretation of significance tests or assessments (e.g. Cox 1958, 2006, pp. 36, 70, 162, 198; Eysenck 1960; Altman 1991, p. 168; Christensen 2005; Hurlbert & Lombardi 2009). Widespread intuitive understanding of this situation probably accounts for the rarity of lopsided tests in the scientific literature.

Lopsided tests may be becoming less rare, however, in biomedical statistics, where, more than in other fields, we have the additional influence of a long-standing and widespread concern over maximum potential set-wise (or family-wise or experiment-wise) type I error rates (SWERs or $\alpha_S$). Moyé (2000, pp. 154–156, 191–204, 2006b, pp. 83–86, 167–180), for example, warns about the inappropriateness of one-tailed tests and advocates instead use of lopsided tests where a larger portion of $\alpha$ is allocated to the tail representing a potential harmful effect of the therapy being tested. This is intended to increase the power for the detection of such harmful effects and thereby help meet the ethical obligation to minimize the likelihood of a harmful therapy being approved for widespread use. Moyé does not advocate any particular allocation ratio, but he presents a number of hypothetical examples. In one of these, $\alpha_C$ is set at 0.125, with 0.10 allocated to the tail representing harm and 0.025 to the tail representing benefit (Moyé 2000, p. 155, 2006b, p. 178).

Unlike the lopsided-test advocates cited earlier in this review who were concerned only with individual significance tests, for Moyé and some other biomedical statisticians involved with clinical trials the allocation of $\alpha$ between tails is only a minor component of a larger and more complex alpha-allocation, alpha-budgeting or alpha-spending paradigm. This paradigm has been developing for half a century, and is now the subject of a very large literature, accessible through, for example, Tukey (1953), Miller (1981), Hochberg & Tamhane (1987), Braun (1994), Moyé (2003, 2006a, 2008), Dmitrienko *et al.*(2003, 2005, 2007) and Dmitrienko & Tamhane (2007). The paradigm combines: (i) the fixing of $\alpha_S$ for an entire experiment or study or, alternatively, for each of a few sets or 'families' of statistical tests within a study; (ii) a rigid distinction between 'exploratory' statistical hypotheses and 'confirmatory' ones; (iii) a rigid dichotomization of 'confirmatory' hypotheses into 'primary' and 'secondary' ones; (iv) a partitioning of $\alpha_S$ among the tests of confirmatory statistical hypotheses, with larger shares allocated to primary than to secondary hypotheses; (v) a lopsided allocation of the small alpha 'fragment' set aside for any given statistical test; and, finally, (vi) characterization of every individual result as 'significant' or 'non-significant', and the whole study as 'positive' or 'negative.'

Moyé (2000, p. 193) gives an example in which there are one 'primary' and three 'secondary' statistical hypotheses to be tested. Out of an overall $\alpha_S$ of 0.10, these hypotheses are allocated individual comparison-wise error rates ($\alpha_C$) of 0.04, 0.035, 0.014 and 0.014, respectively. These $\alpha_C$ values are then lopsidedly allocated to the 'adverse' and 'beneficial' tails of the test statistic distributions in the ratio of 5:3, 3:2, 5:2 and 5:2, for the four tests respectively. In his second edition, that example is replaced by one with three 'primary' and two 'secondary' hypotheses, and he recommends that an overall $\alpha_S$ be partitioned *only among the 'primary' hypotheses* (Moyé 2006b, p. 192). Each 'secondary' hypothesis is tested independently with $\alpha_C = 0.05$, but no matter how low a $P$-value is obtained, no firm ('confirmatory') conclusion is allowed to be based on it.

On top of this complexity of 'hierarchically-ordered multiple objectives' one can also apply so-called 'step-wise' or 'gatekeeping' procedures (e.g. Marcus *et al.* 1976; Westfall & Krishen 2001; Dmitrienko *et al.* 2003, 2005, 2007; Dmitrienko & Tamhane 2007). These can

provide additional power for testing 'primary' simple or composite hypotheses at the expense of perhaps completely foregoing analysis of 'secondary' hypotheses, even if the latter also represent serious questions that the study was designed to answer!

This state of affairs in biomedical statistics reflects a dynamic but unproductive interplay between the objectives of corporations, regulators and statisticians that does not serve the interests of science and medicine well. Pharmaceutical companies understandably want clear, stationary 'goalposts' or requirements that they can tailor their expensive research programs and clinical trials to meet. Regulatory agencies want the decision process to be as mechanical and non-subjective as possible in order to minimize the possibility of their being accused of bad or arbitrary decisions, to avoid battles with any of their constituencies (pharmaceutical companies, researchers, politicians, courts, general public), and to minimize potential confusion caused by turnover in agency personnel. Many statisticians, especially more theoretically inclined ones, will enjoy mathematical complexity and challenging new statistical approaches more than simple data analysis or tutoring of others in such. Thus, it is claimed that the regulatory 'reality is, and is likely to remain, that some sort of per trial (i.e. per experiment) control of the type I error rate is required of sponsors (i.e. research teams and corporations)', regardless of the number of treatment arms, the number of response variables and the number of monitoring dates (Senn & Bretz 2007).

Clinical trials do involve complexities not found in most other types of scientific research. These include the balancing of scientific, commercial and ethical objectives, stopping rules that reflect such balance, scheduling of interim analyses, staggered entry of subjects into trials, the importance of maintaining blinding, the desirability of assessing the influence of different stratification factors (e.g. age, sex, race) when possible, and the need to convince a regulatory agency that the experimental therapy has a definite net positive effect on health. This requires a judicious but ultimately subjective weighing of information on multiple endpoints, namely multiple measures of both potential harm and potential benefit. At a superficial level, the complexity of these needs and objectives might seem well matched by the current fashion for complex statistical procedures in clinical trials. We believe, however, that the additional complexities and obscurities created by subjectively defined sets, $\alpha_S$–values, and alpha-spending algorithms do nothing to serve the legitimate needs and objectives of clinical trials.

If we seem to have focused unduly on Moyé in this section, it is partly because he has been one of the most prolific and enthusiastic proponents of complex alpha-allocation/fixed SWER schemes. These are indeed favoured by many others, including the US Federal Drug Administration. But it is also because in his laudable attempts to be comprehensible he has ended up incorrectly defining key concepts. Primary among these is his definition of $P$ as 'the probability that there is no effect in the population' (Moyé 2000, p. 52), that is, that the null hypothesis is true. He gives that or similar equally misleading definitions of $P$ at many other places in his books (e.g. Moyé 2000, p. 28, 2003, pp. 14, 86, 105, 2006b, p. 87). Similarly, in defining $\alpha_S$ as 'the probability that at least one type I error has occurred across all the analyses' (e.g. Moyé 2003, p. 108, 2006b, p. 187), he errs by omitting the critical qualifier, 'if in fact all null hypotheses are true'. The true probability of a type I error is never known in practice but usually is likely to be $\ll \alpha_S$. It is clear that Moyé himself would not advocate those definitions in discussions with other professional statisticians. One nevertheless wonders how his having articulated them may have influenced subconsciously his strong philosophical preference for SWERPs and complex alpha-allocation schemes.

## 6. Multiplicity is not a problem

This is not the place for a full review of each of the many multiple comparison procedures that employ fixed set-wise type I error rates (SWERPs). However, as these procedures seem to be based on misconceptions similar to those favouring lopsided tests, a brief analysis of the fundamental problems with the SWERP concept is in order. The key problem to which we refer to is not a logical conflict between the premises and assumptions underlying SWERPs and the calculations recommended for carrying them out; rather, it is the logical disconnect between those premises and assumptions and the questions researchers wish to ask and the judicious interpretations of data they hope, or should hope, to achieve. In all basic and applied research, significance tests function mainly to assist assessment of the existence, sign and magnitude of effects, and worry over the maximum probability that one or more type I errors might have been made in some arbitrarily defined set of tests is unwarranted.

In a lengthy, disjointed work published 41 years after its completion, Tukey (1953) forcefully summarized the case for fixing set-wise error rates, which he felt 'should be the standard' (Tukey 1953, p.153). His focus was multiplicities that result from an experiment having multiple (>2) treatments. He discussed how to 'spend' $\alpha_S$, systems of allowances by which portions of an $\alpha_S$ could be allocated to individual comparisons. He labelled as 'a horrible point of view' the idea that one might prefer to stick with comparison-wise error rates ($\alpha_C$ or just *P*-values) (Tukey 1953, p.31) and was to publish several other papers (e.g. Tukey 1977, 1991) written from this viewpoint. In his introduction to Tukey's collected works, Braun (1994) summarizes the influence of Tukey and other promoters of SWERPs by noting, 'By now, most statisticians would agree that the interpretability of a family of statements [or set of tests] requires the overall [set-wise type I] error rate to be bounded at levels ordinarily attached to single statements [i.e. 0.05].' We doubt this is true for most statisticians or scientists, however accurately it may reflect the attitudes of those producing the burgeoning literature on SWERPs and FDRPs. Perry (1986) noted that several British journals and research institutes in the biological sciences explicitly recommend against use of SWERPs. Nickerson (2000, p. 272) suggests that researchers in psychology also mostly refrain from use of SWERPs, while Benjamini *et al.* (2001) state that they are 'mandatory in psychological research' but generally not in medical journals. Curran-Everett (2000) found that SWERPs (other than Fishers's LSD method) were used in only 28 per cent of 798 research reports randomly selected from the 1997 volumes of the nine journals published by the American Physiological Society. Our own experience is that, happily, fewer editors and referees for ecology, psychology and behaviour journals are demanding 'corrections' for multiplicity than was the case 20 years ago.

In 1971, after a large number of SWERPs had been devised and were in use, O'Neill & Wetherill (1971) presented at a meeting of the Royal Statistical Society a review entitled, *The present state of multiple comparison methods*. They started off by noting that 'there is still much confusion as to what the basic problems really are, what the various procedures achieve, and what criteria and properties should be studied', and then proceeded to go through the then available procedures in a rather descriptive way. Extensive comments provided by seven discussants following the paper were published with it, and three seem especially relevant. Professor R. L. Plackett (p. 242) stated, 'The whole area is entwined with rigid stylized thinking, expressed by concepts such as error rates and significant differences. Matters would not be improved by further work with decision-theoretic formulations . . . my view [is] that

much of the subject of multiple comparisons is essentially artificial.' Dr J. A. Nelder, Head of the Statistics Department at Rothamsted Experimental Station, said (p. 244) 'In my view, multiple comparison methods have no place at all in the interpretation of data. Their principal use appears to be to lend an air of respectability to otherwise uninteresting sets of data.' In their response to these and other discussants, O'Neill & Wetherill (1971, p. 249) acknowledged that, 'Clearly, there is a strong feeling that the whole subject of multiple comparison is rather artificial . . . [W]e cannot wholly agree . . . However, we recognize that there is some validity in the point Dr. Nelder makes and in view of this, we doubt if there is much point in being too precise about the methods used.'

Has anything changed to give stronger justification for SWERPs? No. Many books and articles still promote use of fixed set-wise error rates. An *International Conference on Multiple Comparison Procedures* is held every two to three years, the 6th having taken place in March 2009. However, many statisticians and scientists from different disciplines consider that Mssrs Nelder and Plackett were correct in their negative assessments of the appropriateness of SWERPs (e.g. Wilson 1962; Cox 1965; Duncan 1965; Little 1978; Carmer & Walker 1982; Preece 1982; O'Brien 1983; Perry 1986; Finney 1988; Mead 1988 p.311; Hurlbert 1990; Rothman 1990; Keppel 1991; Soto & Hurlbert 1991; Pearce 1993; Savitz & Olshan 1995, 1998; Stewart-Oaten 1995; Pocock 1997; Perneger 1998; Crabbe *et al.* 1999; Hurlbert & Lombardi 2003; Moran 2003; O'Keefe 2003; Nakagawa 2004; Schulz & Grimes 2005). Some very good statistics texts do not bother to cover SWERPs at all (e.g. Mead & Curnow 1983; Freedman *et al.* 1991; Spanos 1999; Box *et al.* 2005) or treat them only briefly and dismissively (e.g. Mead 1988; Mead *et al.* 2003). Such omission is a strong statement in itself, given that almost every study and every publication involves a multiplicity of tests.

Whatever statistical tests are dictated by the objectives and design of a study are best carried out one-by-one without any adjustments for multiplicities, whether these derive from there being multiple treatments, multiple monitoring dates or multiple response variables. Clarity and interpretability of results will be favoured. Hurlbert & Lombardi (2009) cite 12 articles from the biological literature exemplifying the neo-Fisherian approach to significance assessment. Multiple statistical tests of various kinds are carried out in each article. No 'corrections' for multiplicities are employed, and no author specifies an $\alpha_C$ or uses the label 'significant'.

Miller (1981, p. 33) in the introduction to his book on simultaneous inference (SWERPs broadly understood) acknowledges that, 'Provided the nonsimultaneous statistician [i.e. one who does not fix his $\alpha_S$'s] and his client are well aware of their [maximum potential type I] error rates for groups of statements, the author can find no quarrel with them. Every man should get to pick his own error rates.' Most recent reviews of SWERPs, however, are less generous and are premised on the assumption that control of $\alpha_S$ is obligatory (e.g. Day & Quinn 1989 [but see Quinn & Keough 2002]; Hochberg & Tamhane 1987; Shaffer 1995, 2006; Curran-Everett 2000; Moyé 2000, 2007; Bender & Lange 2001; Dmitrienko *et al.* 2005, 2007; Farcomeni 2008).

Miller also acknowledged the lack of any objective way of defining the family (or set) for which $\alpha_S$ is to be fixed. In order to get on with the substance of his book, he simply posited that, 'The *natural family* for the author *in the majority of instances* is the individual experiment of a *single researcher* . . . [although] large single experiments cannot be treated as a whole without an unjustifiable loss of sensitivity' (Miller 1981, p. 34). But what about the other 'instances'? How about long-term experiments in medicine, ecology, hydrology, and other fields where many response variables are monitored over a few to many years and

many manuscripts by different authors are published over time? In any case, as Saville (1990) notes, 'An experiment is no more a natural unit than a project consisting of several experiments or a research program consisting of several projects.' Hochberg & Tamhane (1987), Westfall & Young (1993), Shaffer (1995), Moyé (2003) and Dmitrienko *et al.* (2005) also consider the difficulty of delimiting 'families' and are unable to offer clear guidelines. Some suggest dichotomously partitioning sets of tests according to somewhat arbitrary distinctions of 'exploratory' versus 'confirmatory' and 'primary' versus 'secondary' hypotheses, and, in medicine, of 'beneficial' versus 'harmful' effects. Every such partition, of course, has the 'happy' effect of reducing set size! To our mind, such partitionings represent neither objective nor generalizable procedures. They amount to little more than verbal gymnastics and special pleading of researchers desperate to retain some power (via small set sizes) and not yet bold enough to divorce themselves from the whole notion of fixed SWERs. International guidelines for clinical trials (ICH 1999) state that multiplicities 'may necessitate an adjustment to type I error [but that] . . . methods to avoid or reduce multiplicity are sometimes preferable when available.' Bender & Lange (2001) likewise suggest, 'The easiest and best interpretable approach is to avoid multiplicity as far as possible' in the design of studies. In other words, maximize the *inefficiency* of your research program! Use two treatments even though your objectives would be better served by four. Monitor only one or two response variables at one or two points in time even though you have interest in the detailed time course of twenty.

The SWERP 'cottage industry' (Tukey 1991) seems to have generated large amounts of mathematics divorced from the real needs of researchers, physicians, regulators and the public. The long-standing, strong arguments against fixing set-wise $\alpha_S$ are not rebutted by SWERP enthusiasts – they are mostly just ignored.

These arguments include: (i) the irrelevance of the probability of one or more type I errors for the rare or unrealistic situation in which all nulls in a set being assessed are true; (ii) the dependence of SWERPs on the notion that it is rational to fix alphas for both individual tests and sets of tests in order to generate dichotomized conclusions ('significant'/'non-significant', or 'positive study'/'negative study'); (iii) the lack of any objective grounds for specifying $\alpha_S$, for which reason it is usually blindly set at the familiar 0.05; (iv) the decrease in power of individual comparisons if $\alpha_S$ is set low; (v) the lack of objective grounds for defining the size or scope of a set, as discussed above; (vi) the consequently inconsistent way in which sets are defined in practice; (vii) the penalization of studies designed to answer many questions that results from requiring stronger evidence (lower *P*-values) before their individual null hypotheses can be rejected, as compared with smaller studies using the same $\alpha_S$; (viii) the inconstancy of the evidentiary standard for 'significance' from one test to another within a set, from one set to another within a study, and from one study to another, thus greatly diminishing comparability of analyses and studies (not that 'significant' is a term that should be used to describe *P*-values – see e.g. Cox 1958; Eysenck 1960; Skipper *et al.* 1967; Altman 1991; Hurlbert & Lombardi 2009); and (ix) neglect of the fact that subject matter interpretations logically are made on a test-by-test basis and should be strongly influenced by estimated individual effect sizes, especially in any sort of applied research.

## 7. FDRPs: Old bad wine in a new bottle

The great reduction in power of classical SWERPs as set size increases, especially if $\alpha_S$ is kept at 0.05, has caused most scientists to use the comparison-wise approach and,

increasingly, to present only unadjusted *P*-values in their reports, often even without explicit specification of an $\alpha_C$ (Hurlbert & Lombardi 2009). Other statisticians and scientists who in the past have been rigid in their demand that SWERPs be used are now switching, however, to a slightly less conservative approach to multiple comparisons: false discovery rate procedures (FDRPs).

The true FDR is defined for a set of $m$ statistical tests as the proportion of rejected null hypotheses that represents true nulls; that is, as the ratio of the number ($V$) of actual type I errors or 'false discoveries' to the total number ($R$) of nulls rejected (Benajamini & Hochberg 1995; Benjamini & Yekutieli 2005). The true FDR = $V/R$ for a given set of tests is always unknown, but just as one can define a maximum acceptable potential SWER (without knowing the true SWER) and use it to adjust significance criteria, one can define $q$, a maximum acceptable potential FDR, and use it to adjust significance criteria. This is usually termed 'controlling the FDR'. (Some confusion results from the fact that $q$ is usually referred to as FDR, whereas MFDR, with M standing for 'maximum', would be a clearer way of representing $q$ and what is actually being done. Similarly, MSWER might be recommended.) FDR criteria are a function of the actual *P*-values obtained, including the number ($m$) of them in a set. In the majority of studies, where all or most null hypotheses are likely to be false, $q$ will be much higher than the true FDR, sometimes orders of magnitude higher.

Development of the FDR concept as a way of controlling type I errors began long ago (e.g. Ecklund & Seeger 1965; Seeger 1968; Spjøtvoll 1972; Soric 1989), but modern enthusiasm for it dates mostly from the clear and energetic exposition of the FDRP by Benjamini & Hochberg (1995). That paper has now (February 2012) been cited at least 13,045 times (Google Scholar datum), and the exploding literature on FDRPs is overwhelmingly favourable, with numerous demonstrations of their application in different disciplines. An entry into this literature can be had via Benjamini & Yekutieli (2001, 2005), Verhoeven (2005), Benjamini *et al.* (2006), Genovese *et al.* (2006), Storey (2007), Farcomeni (2008) and Leek & Storey (2008). Curran-Everett (2000) suggests that 'the FDR procedure may be the best practical solution to the problems of multiple comparison that exist within science.' Garcia (2004) states that 'methods based on controlling the false discovery rate (FDR) deserve a more frequent use in ecological studies . . . . [and are] more powerful than the sequential Bonferroni procedures'. Benjamini & Yekutieli (2005) claim that 'False discovery rate control has become an *essential* tool in any study that has a very large multiplicity problem' [our italics]. Unless all the silent statisticians speak up who still advocate, in all or almost all situations, simply reporting individual *P*-values and effect sizes, the day cannot be far off when editors, referees, regulatory agencies and dissertation advisors will be demanding use of FDRPs with the same vigour and self-assuredness as in past decades some of them demanded use of SWERPs.

The new bottle will do nothing for the bad wine of excessive preoccupation with hypothetical set-wise error rates. All FDRPs suffer from the same nine problems, broadly interpreted, as were listed earlier for SWERPs. Those of the listed problems that require it are easily reformulated in terms of $q$ and FDRPs instead of $\alpha_S$ and SWERPs.

The main advantage claimed for FDRPs is that they are more powerful than SWERPs and reduce the probability or frequency of type II error. The claim is, in a sense, based on specious argument. An FDRP and a SWERP applied to the same data set are asking different questions. To compare the results of using a Bonferroni procedure, for example, with $\alpha_S$ set at 0.05 with the results of using a FDRP with $q$ set at 0.05 is to compare apples and oranges. The illusion is given that, in some vague, subjective sense, type I error is fixed at 0.05 in both

cases, so that the greater number of nulls rejected by the FDRP supposedly is demonstrative of its greater power. In fact, of course, a $q = 0.05$ almost always will correspond to an $\alpha_S \gg 0.05$. One might as well compare results for two Bonferroni procedures, one using an $\alpha_S$ of 0.05 and the other using an $\alpha_S$ of 0.15. In going from $\alpha_S = 0.05$ to $q = 0.05$, one is simply relaxing the criteria for 'statistical significance', not gaining power in any substantive, cost-free way.

The flaws of FDRPs are fully displayed in the criticisms that Benjamini & Hochberg (1995) direct at Neuhaus *et al.* (1992), whose data they use to demonstrate the putative usefulness of the FDRP. Neuhaus *et al.* is a report of a randomized clinical trial conducted to compare the effectiveness of two blood clot-dissolving therapies in heart attack victims. In that trial, 41 predominantly categorical response variables (endpoints, in clinical trial jargon) were measured. For all, the event frequencies or means for the two treatments and a *P*-value from a $\chi^2$- or *t*-test are presented in clear, well laid-out tables. Discussion of the results is cogent, focused on effect sizes and their clinical import. All in all, a model report in the best neo-Fisherian tradition (Hurlbert & Lombardi 2009), with not even $\alpha_C$ specified, free (with two minor exceptions, p. 888) of the superfluous 'significant'/'non-significant' jargon, and without any rigid dichotomization of 'primary' and 'secondary' endpoints or of 'confirmatory' and 'exploratory' findings.

Benjamini & Hochberg (1995) start off complaining that Neuhaus *et al.* (1992) give 'no attention to the problem of multiplicity . . . [and] no word of warning regarding their [*P*-values'] interpretation.' They then select a 'family' of 15 of the response variables with which to demonstrate the FDRP and contrast its results with those yielded by the classic Bonferroni procedure. The FDRP involves ranking the *m P*-values from lowest to highest ($i = 1, \ldots, m$), calculating $iq/m$ for each variable tested, and classifying as 'significant' only those differences associated with *P*-values $\leq iq/m$. The results obtained by Benjamini & Hochberg are given in Table 1, along with the original frequencies and *P*-values given by Neuhaus *et al.* (1992) and two indices (effect size, sign odds ratio) calculated by us. The latter provide a useful perspective but contain no information not already given in Neuhaus *et al.* (1992). Other measures of effect size may be preferred by some.

For nine of the 15 response variables, Neuhaus *et al.* (1992) reported a $P \leq 0.05$. Setting $\alpha_S = 0.05$, Benjamini & Hochberg (1995) show that the Bonferroni procedure allows rejection of the null for only three of those nine variables with $P \leq 0.05$. One of the nulls *not* rejected is that for the 3.4-fold greater in-hospital death rate for one treatment over the other ($P = 0.0095$). This is a good demonstration of the low power of Bonferroni methods.

Benjamini & Hochberg (1995) then apply their FDRP to the same set of 15 $\chi^2$-tests, set $q = 0.05$, and find they *now* have grounds for *rejecting* the null of equal in-hospital death rates: the death rates *are* 'significantly different'. It is an unfortunate fact that with all FDRPs, the conclusion about the 'significance' of the difference between in-hospital death rates will be strongly driven not only by the *P*-value for that difference but also by the *P*-values for the other differences, the value of *q* selected, by any increase or decrease in 'family size,' and by the actual *P*-values correspondingly added to or subtracted from the 'family'. For example, if tests on the three top-ranked variables had yielded *P*-values > 0.01, then the difference between treatments in in-hospital death rate would again be found to be 'non-significant'! This is because the FDR criterion ($iq/m$) for that comparison would have dropped from 0.0133 to 0.0033.

**TABLE 1**

*Reanalysis of Benjamini & Hochberg's (1995) application of a FDR procedure to a 'family' of m = 15 response variables out of 41 assessed in a randomized clinical trial (Neuhaus et al. 1992) comparing two procedures for dissolving blood clots in heart attack victims. The number of subjects was 211 for treatment A and 210 for treatment B.*

| Tests ranked by ascending $P$-values ($i$) | Response variable (all categorical, with two states, + and −) | Treatment (A, B)[a] and frequency (%) | | Effect size ($f_A/f_B$) | Sign odds ratio[b] (1-$P$/2)/($P$/2) | $P$-value[c] | Critical value of test criterion | |
|---|---|---|---|---|---|---|---|---|
| | | A | B | | | | FDR ($iq/m$) | Bonferroni ($\alpha_S/m$) |
| 1 | Allergic reaction | 8.6 | 0.5 | 17.2 | 19999 | 0.0001** | 0.0033 | 0.0033 |
| 2 | Bleeding, puncture site | 30 | 15 | 2.00 | 4999 | 0.0004** | 0.0067 | 0.0033 |
| 3 | Bleeding, overall | 45 | 31 | 1.45 | 1052 | 0.0019** | 0.0100 | 0.0033 |
| 4 | In-hospital death, total | 8.1 | 2.4 | 3.37 | 210 | 0.0095* | 0.0133 | 0.0033 |
| 5 | Bleeding, transfusion | 8.1 | 2.8 | 2.89 | 99 | 0.0201 | 0.0167 | 0.0033 |
| 6 | Cardiogenic shock, 48 h | 6.2 | 1.9 | 3.26 | 71 | 0.0278 | 0.0200 | 0.0033 |
| 7 | Blood pressure drop, 90 min | 18 | 10 | 1.80 | 66 | 0.0298 | 0.0233 | 0.0033 |
| 8 | In-hospital death, 48 h | 4.3 | 0.9 | 4.78 | 57 | 0.0344 | 0.0267 | 0.0033 |
| 9 | Cardiogenic shock, 90 min | 1.9 | 0.0 | $\infty$ | 43 | 0.0459 | 0.0300 | 0.0033 |
| 10 | Pericardial tamponade | 1.4 | 0.5 | 2.80 | 5.2 | 0.3240 | 0.0333 | 0.0033 |
| 11 | Blood pressure drop, 48 h | 15 | 12 | 1.25 | 3.7 | 0.4262 | 0.0367 | 0.0033 |
| 12 | Cerebrovascular ischemia | 0.9 | 0.5 | 1.80 | 2.5 | 0.5719 | 0.0400 | 0.0033 |
| 13 | Reinfarction | 4.8 | 3.8 | 1.26 | 2.1 | 0.6528 | 0.0433 | 0.0033 |
| 14 | Recurrent ischemia | 2.9 | 3.3 | 0.88 | 1.6 | 0.7590 | 0.0467 | 0.0033 |
| 15 | Bleeding, cerebral | 0.9 | 0.9 | 1.00 | 1.0 | 1.0000 | 0.0500 | 0.0033 |

[a] Treatments A and B correspond to treatments 'APSAC' and 'rtPA', respectively, in Neuhaus *et al.* (1992).
[b] This is the odds that the sign of an observed difference, $d$, is the same as the sign of the true difference, $\partial$, given the assumption that a true difference exists (Hurlbert & Lombardi 2009).
[c] A single asterisk (∗) indicates that the observed difference is 'statistically significant' by the FDR procedure (with $q = 0.05$); a double asterisk (∗∗) indicates that the difference is 'statistically significant' by both the FDR procedure (with $q = 0.05$) *and* the Bonferroni procedure (with $\alpha_S = 0.05$). All $P$-values are from $\chi^2$-tests on 2 × 2 contingency tables.

The FDRP applied by Benjamini & Hochberg (1995) leaves unrejected the nulls for five response variables where $\chi^2$-tests yield $0.02 < P < 0.05$. How are these to be interpreted? Merely as 'non-significant', 'suggestive' or 'exploratory' findings?

Regardless of how high the maximum potential $\alpha_S$ might be, those five $P$-values are *strong* evidence of *real* treatment effects on these variables. A good clinician will accept that, consider the estimated effect magnitudes and their medical implications, and act accordingly.

Nothing in the post-1995 articles we have read on FDRPs or their modifications (e.g. Dudoit *et al.* 2003; Storey 2003, 2007; Storey *et al.*, 2004, 2007; Fernando *et al.* 2004;

Benjamini & Yekutieli 2005; Cui *et al.* 2005; Benjamini *et al.* 2006; Genovese *et al.* 2006; Storey *et al.* 2007; Farcomeni 2008; Leek & Storey 2008) offers a glimmer of hope that these procedures in all their current variety can be rescued from flawed core assumptions any more than a half-century's efforts have been able to justify SWERPs to most scientists. In studies such as Neuhaus *et al.* (1992), fixing of set-wise error rates of any sort inevitably leads to weak argument, bad science, and poor decision-making.

Are there parallel mismatches between scientific question and statistical procedure in DNA microarray or neuroimaging studies, where often thousands of tests are conducted in a given study and where FDR and related procedures have become very popular (e.g. Genovese *et al.* 2002; Dudoit *et al.* 2003; Benjamini & Yuketieli 2005; Singh & Dan 2006; Storey *et al.* 2007)? Are the questions in these types of studies so different from those in other types of research that completely new methods are needed when we go from five or a hundred tests to thousands? We think not, having seen in wide reading no convincing arguments in support of that idea. We encourage those who think otherwise to take up the challenge!

## 8. Conclusions

SWERPs and FDRPs collectively represent historically understandable but logically unjustifiable extensions of the outdated paleo-Fisherian and Neyman–Pearsonian frameworks for significance testing (see review by Hurlbert & Lombardi 2009). They have been driven by excessive concern over high estimates, when many tests are carried out, of the maximum potential probability ($\alpha_S$) of one or more type I errors or the maximum potential false discovery rate ($q$). Ignored is the fact that in most studies the true probability of type I errors actually occurring is zero or very small, as most nil null hypotheses posited can be expected, on first principles or prior knowledge, to be false. As Schulz & Grimes (2005) state, 'statistical adjustments for multiplicity provide crude answers to an irrelevant question.' What we truly need to guard against are type II errors and type III errors. The first are risked if one naively decides in favour of the null when $P$-values are high. A type III error is defined as concluding that there is an effect in one direction when in fact it is in the opposite direction (e.g. Kaiser 1960; Shaffer 1972, 2006; Schulman 1992). It is especially risked if, when $P$-values are not low, one not only decides against the standard nil null but also concludes that the sign of $\partial$ is the same as the sign of $d$.

Ziliak & McCloskey (2008) recently published an interesting book entitled '*The Cult of Statistical Significance: How the Standard Error Cost Us Jobs, Justice and Lives'*. It expands on those authors' earlier critiques and is a vehemently argued diatribe against R. A. Fisher and significance testing in general. Like much of the criticism of significance tests over the last half-century, the book blames significance tests themselves for their misuse and misinterpretation by scientists and statisticians. The book is nevertheless worth a read, especially if tempered by perusal of some the strong criticisms of it that have already appeared (Hoover & Siegler 2008; Spanos 2008; Hurlbert & Lombardi 2009). The abundant errors and hyperbolic language in Ziliak & McCloskey (2008) are especially unfortunate in that they distract attention from the two most important and worthwhile messages in the book. The first is that, for many researchers, a narrow focus on $P$-values has caused neglect of the important matter of effect sizes and their interpretation. The second big message is that, as recognized by many scientists before them, the fixing of alphas and the dichotomization of all results into 'significant' and 'non-significant' findings has no rational basis despite having served

for almost a century as the basis for conventional paleo-Fisherian and Neyman–Pearsonian approaches to significance testing. It is that dichotomization that has driven much of the inappropriate use of one-tailed tests (Lombardi & Hurlbert 2009), lopsided tests, SWERPs and FDRPs, as discussed in this article. Few are more deserving of Ziliak & McCloskey's (2008) critique than are the proponents of SWERPs and FDRPs. Respite can be provided by neo-Fisherian significance assessment (NFSA: Hurlbert & Lombardi 2009) as introduced at the beginning of this article. In that framework, no CWERs, SWERs or FDRs of any sort can be justified, and the word 'significant' is not used.

Late during revision of this paper, Michael Riggs called our attention to a recent issue of the *Journal of Biopharmaceutical Statistics* that presents an article entitled, *Some controversial multiple testing problems in regulatory applications* (Hung & Wang 2009), plus extended commentaries on that article by five sets of statisticians. Collectively these papers recount many of the long-known difficulties of SWERPs, especially as exacerbated by the dictates of the US Federal Drug Administration. Not surprisingly, no consensus is reached, the concept of fixed SWERs is not challenged, and no desire to adopt a neo-Fisherian framework is expressed. Confusion on this topic clearly remains widespread.

It has been common for SWERP and FDRP enthusiasts to refer to the 'multiplicity problem' as one of the most difficult and severe problems in data analysis. We and many others see no problem with large numbers of tests being conducted and reported independently in a given study and believe that the only serious statistical 'multiplicity problems' are those that are produced by SWERPs and FDRPs themselves – or, occasionally, by dishonest, careless or incomplete reporting.

## 9. Epilogue

Is there anything new under the sun? Perhaps not.

> *The basic question is, what is the most meaningful unit in which to evaluate research? Traditional practice apparently has chosen the [statistical] hypothesis as the unit and this paper maintains that this is the correct choice . . . . It seems foolish for researchers to accept additional statistical complications unless there are telling reasons for doing so. . . . .[I]t is practically impossible for a statistically naïve researcher to abandon the traditional per-hypothesis techniques because statisticians have not yet agreed upon any other strategy or even on how best to achieve the various alternatives that have been advanced. (Wilson 1962)*

> *As with all subjects involving mathematics . . . an ever-increasing body of theory is constructed, and because the developments of the subject [statistics] are largely in the hands of mathematicians there is a tendency to revise and elaborate this theory, sometimes to unnecessary length . . . From an early age, mathematicians are encouraged not to take responsibility for the validity of their assumptions, or for their relevance to the problem they are attempting to solve. (Yates & Healy 1964)*

## *References*

ABELSON, R.P. (1995). *Statistics as Principled Argument*. Hillsdale, NJ: Lawrence Erlbaum.

ALTMAN, D.G. (1991). *Practical Statistics for Medical Research*. New York: Chapman and Hall.

BART, J., FLIGNER, M.A. & NOTZ, W.I. (1998). *Sampling and Statistical Methods for Behavioral Ecologists*. Cambridge: Cambridge University Press.

BENDER, R. & LANGE, S. (2001). Adjusting for multiple testing – when and how? *J. Clin. Epidemiol.* **54**, 343–349.

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300.

BENJAMINI, Y. & YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals Statist.* **29**, 1165–1188.

BENJAMINI, Y. & YEKUTIELI, D. (2005). Quantitative trait loci analysis using the false discovery rate. *Genetics* **171**, 783–790.

BENJAMINI, Y., DRAI, D., ELMER, G., KAFKAFI, N. & GOLANI, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* **125**, 279–284.

BENJAMINI, Y., KRIEGER, A.M. & YEKUTIELI, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507.

BOX, G.E.P., HUNTER, W.G. & HUNTER, J.S. (2005). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, 2nd edn. New York: Wiley.

BRAUN, H.I., ed. (1994). *The Collected Works of John W. Tukey, Volume VIII, Multiple Comparisons: 1948–1983*. New York: Chapman and Hall.

BRAVER, S.L. (1975). On splitting the tails unequally: a new perspective on one- versus two-tailed tests. *Educ. Psychol. Meas.* **35**, 283–301.

CARMER, S.G. & WALKER, W.M. (1982). Baby Bear's dilemma: A statistical tale. *Agron. J.* **74**, 122–124.

CHRISTENSEN, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *Amer. Statist.* **59**, 121–126.

COWLES, M. (1989). *Statistics in Psychology: An Historical Perspective*. Hillsdale, NJ: Lawrence Erlbaum.

COX, D.R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357–372.

COX, D.R. (1965). A remark on multiple comparison methods. *Technometrics* **7**, 223–224.

COX, D.R. (2006). *Principles of Statistical Inference*. Cambridge: Cambridge University Press.

CRABBE, J.C., WAHLSTEN, D. & DUDEK, B.C. (1999). Genetics of mouse behavior: interactions with laboratory environment. *Science* **284**, 1670–1672.

CUI, X., HWANG, J.T., QIU, J., BLADES, N.J. & CHURCHILL, G.A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* **8**, 414–432.

CURRAN-EVERETT, D. (2000). Multiple comparisons: Philosophies and illustrations. *Amer. J. Physiol.: Regul., Integr. Comp. Physiol.* **279**, R1–R8.

DARLINGTON, R.B. & CARLSON, P.M. (1987). *Behavioral Statistics*. New York: The Free Press.

DAY, R.W. & QUINN, G.P. (1989). Comparison of treatments after an analysis of variance in ecology. *Ecol. Monogr.* **59**, 433–463.

DMITRIENKO, A. & TAMHANE, A.C. (2007). Gatekeeping procedures with clinical trial applications. *Pharm. Stat.* **6**, 171–180.

DMITRIENKO, A., OFFEN, W.W. & WESTFALL, P.H. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Stat. Med.* **22**, 2387–2400.

DMITRIENKO, A., MOLENBERGHS, G., CHUANG-STEIN, C. & OFFEN, W. (2005). *Analysis of Clinical Trials Using SAS: A Practical Guide*. Cary, NC: SAS Institute.

DMITRIENKO, A., WIENS, B.L., TAMHANE, A.C. & WANG, X. (2007). Tree-structured gatekeeping tests in clinical trials with hierarchically ordered objectives. *Stat. Med.* **26**, 2465–2478.

DUDOIT, D., SHAFFER, J.P. & BOLDRICK, J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18**, 71–103.

DUNCAN, D.B. (1965). A Bayesian approach to multiple comparisons. *Technometrics* **7**, 171–222.

ECKLUND, G. & SEEGER, P. (1965). Massignifikansanalys. *Statistisk Tidskrift Stockholm, 3d series* **4**, 355–365.

EYSENCK, H.J. (1960). The concept of statistical significance and the controversy about one-tailed tests. *Psych. Rev.* **67**, 269–271.

FARCOMENI, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat. Methods Med. Res.* **17**, 347–388.

FERNANDO, R.L., NETTLETON, D., SOUTHEY, R.R., DEKKERS, J.C.M., ROTHSCHILD, M.F. & SOLLER, M. (2004). Controlling the proportion of false positives in multiple dependent tests. *Genetics* **166**, 611–619.

FINNEY, D.J. (1988). Was this in your statistics textbook? III. Design and analysis. *Exp. Agric.* **24**, 421–432.

FLEISS, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd edn. New York: Wiley.

FLEISS, J.L. (1986). *The Design and Analysis of Clinical Experiments*. New York: Wiley.

FREEDMAN, D., PISANI, R., PURVES, R. & ADHIBARI, A. (1991). *Statistics*, 2nd edn. New York: Norton.

GAINES, S.D. & RICE, W.R. (1990). Analysis of biological data when there are ordered expectations. *Amer. Nat.* **135**, 310–317.

GARCIA, L.V. (2004). Escaping the Bonferroni iron claw in ecological studies. *Oikos* **105**, 657–663.

GENOVESE, C.R., LAZAR, N.A. & NICHOLS, T. (2002). Threshholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* **15**, 870–878.

GENOVESE, C.R., ROEDER, K. & WASSERMAN, L. (2006). False discovery control with p-value weighting. *Biometrika* **93**, 509–524.

GLASS G.V. & HOPKINS, K.D. (1984). *Statistical Methods in Education and Psychology*, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.

GOLDFRIED, M.R. (1959). One-tailed tests and 'unexpected' results. *Psych. Rev.* **66**, 79–80.

HARNETT, D.L. & SONI, A.K. (1991). *Statistical Methods for Business and Economics*, 4th edn. New York: Addison-Wesley.

HARRIS R.J. (2005a). Classical statistical inference extended: split-tailed tests. *Encyclopedia of Statistics in Behavioral Science*, *vol. 1*, eds. B.S. Everitt and D.C. Howell. pp. 263–268. Chichester: Wiley.

HARRIS R.J. (2005b). Classical statistical inference: practice versus presentation. *Encyclopedia of Statistics in Behavioral Science*, *vol. 1*, eds. B.S. Everitt and D.C. Howell. pp. 268–278. Chichester: Wiley.

HARRIS, R.J. & QUADE, D. (1992). The minimally important difference significant criterion for sample size. *J. Educ. Behav. Stat.* **17**, 27–49.

HAWKINS, D. (2005). *Biomeasurement*. New York: Oxford University Press.

HOCHBERG, Y. & TAMHANE, A.C. (1987). *Multiple Comparison Procedures*. New York: Wiley.

HOOVER, D.K. & SIEGLER, M.V. (2008). Sound and fury: McCloskey and significance testing in economics. *J. Econ. Methodol.* **15**, 1–37.

HUNG, H.M.J. & WANG, S.-J. (2009). Some controversial multiple testing problems in regulatory applications. *J. Biopharm. Statist.* **19**, 1–11.

HURLBERT, S.H. (1990). Pastor binocularis: now we have no excuse [Review of *The Design of Experiments*, by R. Mead]. *Ecology* **71**, 1222–1223.

HURLBERT, S.H. & LOMBARDI, C.M. (2003). Design and analysis: uncertain intent, uncertain result [Review of *Experimental Design and Data Analysis for Biologists*, by G.P. Quinn & M.J. Keough]. *Ecology* **83**, 810–812.

HURLBERT, S.H. & LOMBARDI, C.M. (2009). Final collapse of the Neyman-Pearson decision-theoretic framework and rise of the neoFisherian. *Ann. Zool. Fenn.* **46**, 311–349.

ICH (1999). ICH harmonised tripartite guideline: statistical principles for clinical trials. E9 Expert Working Group, International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. *Stat. Med.* **18**, 1905–1942

KAISER, H.F. (1960). Directional statistical decisions. *Psych. Rev.* **67**, 160–167.

KENDALL, M. & STUART, A. (1979). *The Advanced Theory of Statistics*, 2, *Inference and Relationship*, 4th edn. London: Griffin.

KEPPEL, G. (1991). *Design and Analysis: A Researcher's Handbook*, 3rd edn. Englewood Cliffs, NJ: Prentice Hall.

KIMMEL, H.D. (1957). Three criteria for the use of one-tailed tests. *Psych. Bull.* **16**, 345–353.

KIRK, R.E. (1982). *Experimental Design*, 2nd edn. Pacific Grove, CA: Brooks/Cole Publishing.

KLINE, R.B. (2004). *Beyond Significance Testing*. Washington, DC: American Psychological Association.

KORNILOV, S.G. (1993). Errors in the description of the $F$ test and some thoughts on one-sided statistical tests. *Industr. Lab.* **59**, 720–725. (*Translation of Russian article published in: Zavodskaya Laboratoriya* **59**, 60, 1993).

LEEK, J.T. & STOREY, J.D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105**, 18718–18723.

LITTLE, T.M. (1978). If Galileo published in *HortScience*. *HortScience* **13**, 504–506.

LOMBARDI, C.M. & HURLBERT, S.H. (2009). Misprescription and misuse of one-tailed tests. *Austral Ecol.* **34**, 447–468.

MANTEL, N. (1983). Ordered alternatives and the 1–1/2 tail test. *Amer. Statist.* **37**, 225–228.

MARCUS, R., PERITZ, E. & GABRIEL, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

MARTIN, P. & BATESON, P. (2007). *Measuring Behaviour: An Introductory Guide*, 3rd edn. Cambridge: Cambridge University Press.

MEAD, R. (1988). *The Design of Experiments*. Cambridge: Cambridge University Press.

MEAD R. & CURNOW, R.N. (1983). *Statistical Methods in Agriculture and Experimental Biology*. New York: Chapman and Hall.

MEAD, R., CURNOW, R.N. & HASTED, A.M. (2003). *Statistical Methods in Agriculture and Experimental Biology, 3rd edn*. New York: Chapman & Hall.

MEEK, G.E. & OZGUR, C.O. (2004). Unequal division of type I risk in statistical inferences. *Decision Sci. J. Innov. Educ.* **2**, 45–55.

MEEK, G.E. & TURNER, S.J. (1983). *Statistical Analysis for Business Decisions*. Upper Saddle River, NJ: Houghton and Mifflin.

MILLER, R.G., JR. (1981). *Simultaneous Statistical Inference*, 2nd edn. New York: Springer.

MOORE, D.S. & MCCABE, G.P. (1989). *Introduction to the Practice of Statistics*. New York: W.H. Freeman.

MORAN, M.D. (2003). Arguments for rejecting the sequential Bonferroni in ecological studies. *Oikos* **100**, 403–405.

MOYÉ, L.A. (2000). *Statistical Reasoning in Medicine: The Intuitive P-value Primer*. New York: Springer.

MOYÉ, L.A. (2003). *Multiple Analyses in Clinical Trials: Fundamentals for Investigators*. New York: Springer.

MOYÉ, L.A. (2006a). *Statistical Monitoring of Clinical Trials*. New York: Springer.

MOYÉ, L.A. (2006b). *Statistical Reasoning in Medicine: The Intuitive P-value Primer*, 2nd edn. New York: Springer.

MOYÉ, L.A. (2008). Disciplined analyses in clinical trials: The dark heart of the matter. *Stat. Meth. Med. Res.* **17**, 253–264.

NAKAGAWA, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behav. Ecol.* **15**, 1044–1045.

NEALE, M.C. & MILLER, M.B. (1996). The use of likelihood-based confidence intervals in genetic models. *Behav. Genet.* **27**, 113–120.

NEUHAUS, K.-L., VON ESSEN, R., TEBBE, U., VOGT, A., ROTH, M., REISS, M., NIEDERER, W., FORYCKI, F., WIRTZFELD, A., MAEURER, W., LIMBOURG, P., MERX, W. & HAERTEN, K. (1992). Improved thrombolysis in acute myocardial infarction with front-loaded administration of Alteplase: results of the rt-PA–APSAC patency study (TAPS). *J. Amer. Coll. Cardiol.* **19**, 885–891.

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

NOSANCHUK, T.A. (1978). Serendipity tails: a note on two tailed hypothesis tests with asymmetric regions of rejection. *Acta Sociol.*, **21**, 249–253.

OAKES, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. New York: Wiley.

O'BRIEN, P.C. (1983). The appropriateness of analysis of variance and multiple-comparison procedures. *Biometrics* **93**, 787–788.

O'KEEFE, D.J. (2003). Colloquy: Should familywise alpha be adjusted? Against familywise alpha adjustment. *Human Commun. Res.* **29**, 431–447.

O'NEILL, R. & WETHERILL, G.B. (1971). The present state of multiple comparison methods. *J. Roy. Stat. Soc. Ser. B Stat. Methodol.* **33**, 218–250.

PEARCE, S.C. (1993). Data analysis in agricultural experimentation. III. Multiple comparisons. *Exper. Agric.* **29**, 1–8.

PERNEGER, T.V. (1998). What's wrong with Bonferroni adjustments. *Brit. Med. J.* **316**, 1236–1238.

PERRY, J.N. (1986). Multiple-comparison procedures: a dissenting view. *J. Econ. Entomol.* **79**, 1149–1155.

PILLEMER, D.B. (1991). One- versus two-tailed hypothesis tests in contemporary educational research. *Educ. Researcher* **20**(9), 13–17.

POCOCK, S.J. (1997). Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Controll. Clin. Trials* **18**, 530–545.

PREECE, D.A. (1982). The design and analysis of experiments: What has gone wrong? *Util. Math.* **21A**, 210–244.

QUINN, G.P. & KEOUGH, M.J. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge: Cambridge University Press.

RAMSEY, P.H. (1990). 'One-and-a-half-tailed' tests of significance. *Psych. Reports* **66**, 653–654.

RICE, W.R. & GAINES, S.D. (1994a). Extending nondirectional heterogeneity tests to evaluate simple ordered alternative hypotheses. *Proc. Natl. Acad. Sci. USA* **91**, 225–226.

RICE, W.R. & GAINES, S.D. (1994b). The ordered heterogeneity family of tests. *Biometrics* **50**, 746–752.

RICE, W.R. & GAINES, S.D. (1994c). 'Heads I win, tails you lose': Testing directional alternative hypotheses in ecological and evolutionary research. *Trends Ecol. Evol.* **9**, 235–237.

ROTHMAN, K.J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology* **1**, 43–46.

SAVILLE, D.J. (1990). Multiple comparison procedures: the practical solution. *Amer. Statist.* **44**, 174–180.

SAVITZ, D.A. & OLSHAN, A.F. (1995). Multiple comparisons and related issues in the interpretation of epidemiologic data. *Am. J. Epidemiol.* **142**, 904–908.

SAVITZ, D.A. & OLSHAN, A.F. (1998). Describing data requires no adjustment for multiple comparisons: a reply from Savitz and Olshan. *Am. J. Epidemiol.* **147**, 813.

SCHULMAN, R.S. (1992). *Statistics in Plain English*. New York: Van Nostrand Reinhold.

SCHULZ, K.F. & GRIMES, D.A. (2005). Multiplicity in randomized trials I: Endpoints and treatments. *Lancet* **365**, 1591–1595.

SEEGER, P. (1968). A note on method for the analysis of significances en masse. *Technometrics* **10**, 586–593.

SENN, S. & BRETZ, F. (2007). Power and sample size when multiple endpoints are considered. *Pharm. Statist.* **6**, 161–170.

SHAFFER, J.P. (1972). Directional statistical hypotheses and comparisons among means. *Psych. Bull.* **77**, 195–197.

SHAFFER, J.P. (1995). Multiple hypothesis testing. *Ann. Rev. Psychol.* **46**, 561–584.

SHAFFER, J.P. (2006). Recent developments towards optimality in multiple hypothesis testing. *IMS Lecture Notes–Monograph Series* **49**, 16–32.

SIEGEL, S.,(1956). *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

SIEGEL, S. & CASTELLAN, N.J., Jr. (1988) *Nonparametric Statistics for the Behavioral Sciences*, 2d edn. New York: McGraw-Hill.

SINGH, A. & DAN, I. (2006). Exploring the false discovery rate in multichannel NIRS. *Neuroimage* **33**, 542–549.

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. *Amer. Sociol.* **2**, 16–18.

SNEDECOR, G.W. & COCHRAN, W.G. (1989). *Statistical Methods*, 8th edn. Ames, IO: Iowa State University Press.

SOKAL, R.R. & ROHLF, F.J. (1995). *Biometry*, 3rd edn. San Francisco: Freeman.

SORIC, B. (1989). Statistical "discoveries" and effect-size estimation. *J. Amer. Statist. Assoc.* **84**, 608–610.

SOTO, D. & HURLBERT, S.H. (1991). Long-term experiments on calanoid-cyclopoid interactions. *Ecol. Monogr.* **61**, 245–265.

SPANOS, A. (1999). *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge: Cambridge University Press.

SPANOS, A. (2008). Review of *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*, by S.T. Ziliak and D.N. McCloskey. *Erasmus J. Philos. Econ.* **1**, 154–164.

SPJØTVOLL, E. (1972). On the optimality of some multiple comparison procedures. *Ann. Math. Statist.* **43**, 398–411.

STEWART-OATEN, A. (1995). Rules and judgments in statistics: three examples. *Ecology* **76**, 2001–2009.

STOREY, J.D. (2003). The positive false discovery rate: A Bayesian interpretation and the $q$-value. *Ann. Statist.* **6**, 2013–2035.

STOREY, J.D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69**, 347–368.

STOREY, J.D., TAYLOR, J.E. & SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66**, 187–205.

STOREY, J.D., DAI, J.Y. & LEEK, J.T. (2007). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics* **8**, 414–432.

TUKEY, J.W. (1953). The problem of multiple comparisons. *The Collected Works of John W. Tukey, Volume III*, ed. H.I. Braun, pp. 1–300 [1994]. New York: Chapman and Hall. [Work completed and privately circulated starting in 1953].

TUKEY, J.W. (1977). Some thoughts on clinical trials, especially problems of multiplicity. *Science*, **198**, 679–684.

TUKEY, J.W. (1991). The philosophy of multiple comparisons. *Stat. Sci.* **6**, 100–116.

UNDERWOOD, A.J. (1997). *Experiments in Ecology.* London: Blackwell.

VERHOEVEN, K.J.F. (2005). Implementing false discovery rate control: increasing your power. *Oikos* **108**, 643–647.

WELKOWITZ, J., EWEN, R.B. & COHEN, J. (1971, 1991, 1999). *Introductory Statistics for the Behavioral Sciences*, 1st, 4th, 5th edns. New York: Harcourt Brace Jovanovich.

WELKOWITZ, J., COHEN, B.H. & EWEN, R.B. (2006). *Introductory Statistics for the Behavioral Sciences*, 6th edn. New York: Wiley.

WESTFALL, P.H. & KRISHEN, A. (2001). Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *J. Statist. Plann. Inference* **99**, 25–40.

WESTFALL, P.H. & YOUNG, S.S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*. New York: Wiley.

WILSON, W. (1962). A note on the inconsistency inherent in the necessity to perform multiple comparisons. *Psych. Bull.* **59**, 296–300.

YATES, R. & HEALY, M.J.R. (1964). How should we reform the teaching of statistics? *J. Roy. Statist. Soc. Ser. A* **127**, 199–210.

ZAR, J.H. (2004) *Biostatistical Analysis*, 4th edn. New York: Prentice-Hall, Inc.

ZILIAK, S.T. & MCCLOSKEY, D.N. (2008). *The Cult of Statistical Significance: How the Standard Error Cost Us Jobs, Justice and Lives*. Ann Arbor, MI: University of Michigan Press.