If Statistical Significance Tests Are Broken/Misused, What Practices Should Supplement or Replace Them?

Bruce Thompson
TEXAS A&M UNIVERSITY AND BAYLOR COLLEGE OF MEDICINE

ABSTRACT. Given some consensus that statistical significance tests are broken, misused or at least have somewhat limited utility, the focus of discussion within the field ought to move beyond additional bashing of statistical significance tests, and toward more constructive suggestions for improved practice. Five suggestions for improved practice are recommended; these involve (a) required reporting of effect sizes, (b) reporting of effect sizes in an interpretable manner, (c) explicating the values that bear upon results, (d) providing evidence of result replicability, and (e) reporting confidence intervals. Though the five recommendations can be followed even if statistical significance tests are reported, social science will proceed most rapidly when research becomes the search for replicable effects noteworthy in magnitude in the context of both the inquiry and personal or social values.

KEY WORDS: effect size, hypothesis tests, null hypotheses, significance tests, statistical significance

A few years ago Pedhazur and Schmelkin (1991) asserted that 'probably very few methodological issues have generated as much controversy' (p. 198) as have the use and interpretation of statistical significance tests. These tests have certainly proven surprisingly resistant to repeated efforts 'to exorcise the null hypothesis' (Cronbach, 1975, p. 124). Particularly noteworthy among the historical efforts to accomplish the exorcism have been works by Rozeboom (1960), Morrison and Henkel (1970), Carver (1978), Meehl (1978), Shaver (1985) and Oakes (1986). The entire Volume 61, Number 4 issue of the *Journal of Experimental Education* (1993) was devoted to these themes. Yet, notwithstanding the long-term availability of these publications, even today some psychologists still do not understand what statistical significance tests do and do not do.

In a public-domain brief digest disseminated as a class handout by the US Department of Education Educational Resources Information Center, the

THEORY & PSYCHOLOGY Copyright © 1999 SAGE Publications. VOL. 9(2): 165–181
[0959-3543(199904)9:2;165–181;006933]

present author (Thompson, 1994a) provided some simple tests of understanding of what $p_{\text{CALCULATED}}$ actually evaluates:

> In which one of each of the following [three] pairs of studies will the $p_{\text{CALCULATED}}$ be smaller?
>
> —In two studies each involving three groups of subjects each of size 30, in one study the means were 100, 100, and 90, and in the second study the means were 100, 100, and 100.
>
> —In two studies each comparing the standard deviations (*SD*) of scores on the dependent variable of two groups of subjects, in both studies $SD_1 = 4$ and $SD_2 = 3$, but in study one the sample sizes were 100 and 100, while in study two the samples sizes were 50 and 50.
>
> —In two studies involving a multiple regression prediction of *Y* using predictors X_1, X_2, and X_3, and both with samples sizes of 75, in study one $R^2 = .49$ and in study two $R^2 = .25$. (p. 5)

These judgments do not require calculations or additional information. However, making such judgments does require a genuine understanding of what statistical significance tests are all about.[1]

It is not clear how well most authors of journal articles would do on the previous three-item evaluation (cf. Falk & Greenbaum, 1995; Nelson, Rosenthal, & Rosnow, 1986; Oakes, 1986; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). Many of us continue to prefer ‘investing . . . [these tests] with what appear to be magical powers’ (Pedhazur & Schmelkin, 1991, p. 198). And some of us try to use *p* values to cling to a mantle of unattainable objectivity.

The use of statistical tests has recently stimulated yet more controversy. Harlow, Mulaik and Steiger (1997) provide a compendium of views on these issues (for a review, see Thompson, 1998a). Contemporary commentaries include those provided by Hunter (1997), Kirk (1996), Schmidt (1996) and the present author (Thompson, 1996, 1997). The less positive treatments of statistical significance tests have also provoked reactions from test advocates (cf. Chow, 1988; Frick, 1996; Greenwald, Gonzalez, Harris, & Guthrie, 1996; Hagen, 1997; Robinson & Levin, 1997). Yet even Frick (1996) acknowledged that critics of conventional practices ‘usefully point out the limitations of null hypothesis testing’ (p. 388).

Given growing consciousness regarding these limitations, the APA Board of Scientific Affairs recently named a Task Force on Statistical Inference (Shea, 1996). The APA Task Force is charged with recommending policies and practices leading to more informed and thoughtful statistical analyses, including those involving the use of statistical significance tests.

Articles within the *American Psychologist*, published on a seemingly periodic basis, have especially informed the movement of the field as regards statistical significance testing. Table 1 lists some of these articles, and also reports citation frequencies for the articles as of 1996. These *American Psychologist* articles, and the related comments published within

TABLE 1. Citations of selected *American Psychologist* articles

		Number of citations							
Year	Author(s)	Pre-1991	1991	1992	1993	1994	1995	1996*	Total
1994	Cohen	–	–	–	–	–	17	42	59
1991	Rosenthal	–	–	6	2	2	2	1	13
1990	Cohen	–	18	36	38	23	28	23	166
1989	Rosnow & Rosenthal	4	18	23	20	14	12	13	104
1988	Kupfersmid	19	3	6	6	2	4	1	41
1987	Dar	10	2	6	2	2	1	1	24

* The most current Index at the time of this compilation only covered 1996 through September of that year.

the journal, have considerably influenced psychology and the social sciences more generally. For example, Roger Kirk (1996) characterized the two *American Psychologist* articles by Cohen as ‘classics’, and argued that ‘the one individual most responsible for bringing the shortcomings of hypothesis testing to the attention of behavioral and educational researchers is Jacob Cohen’ (p. 747).

The present paper briefly reviews some of the consensus that has arisen or seems to be occurring as regards the use and limits of statistical significance tests. However, the present treatment also explores both (a) recommendations involving changes in research practices and editorial policies and (b) related issues that the field has yet to resolve. Given some consensus that statistical significance tests are broken, misused or at least have somewhat limited utility, the focus of discussion within the field ought to move beyond additional bashing of statistical significance tests, and toward more constructive suggestions for improved practice.

Emerging Consensus

The field appears to have achieved or is approaching consensus regarding certain limitations of statistical significance tests, notwithstanding some psychological resistance (Schmidt & Hunter, 1997; Thompson, 1998b). At least three noteworthy realizations can be briefly cited.

Result Effect Size

First, researchers have recognized that *p values are not useful as indices of study effect sizes* (although some researchers still may implicitly deem more important those studies reporting smaller *p* values—cf. Rosenthal & Gaito, 1963; Zuckerman et al., 1993). The calculated *p* values in a given study are a function of several study features, but are particularly influenced by the confounded, joint influence of study sample size and study effect sizes. Because *p* values are confounded indices, in theory 100 studies with varying sample sizes and 100 different effect sizes could each have the same single $p_{\text{CALCULATED}}$, and 100 studies with the same single effect size could each have 100 different values for $p_{\text{CALCULATED}}$.

This realization led to an important change in the fourth edition of the American Psychological Association Publication Manual (APA, 1994). The manual noted that

> Neither of the two types of probability values [statistical significance tests] reflects the importance or magnitude of an effect because both depend on sample size. . . . You are [therefore] *encouraged* to provide effect size information. (APA, 1994, p. 18; emphasis added)

Result Importance

Second, more and more researchers and editors have come to recognize that *p values do not evaluate result importance*. Therefore, *p* values cannot be used as an effective vehicle for escaping disagreement and confrontation regarding our subjective judgments of the worth of our results. As Thompson (1993) noted,

> . . . importance is a question of human values, and math cannot be employed as an atavistic escape (*à la* Fromm's *Escape from Freedom*) from the existential human responsibility for making value judgments. If the computer package did not ask you your values prior to its analysis, it could not have considered your value system in calculating *p*'s, and so *p*'s cannot be blithely used to infer the value of research results. (p. 365)

Result Replicability

Third, researchers have recognized that $p_{\text{CALCULATED}}$ *values are not informative regarding the likelihood of result replication in future samples* (Thompson, 1996). As Cohen (1994) made so clear, these calculations presume that the null hypothesis exactly describes the population, and then indicate the probability of the sample results (or of sample results even more disparate from the null than those in the actual sample), given the sample size.

But what we want to know is the population parameters, given the statistics in the sample and the sample size. This interest in true population values stems from a desire to avoid the discovery of cold fusion, which leads to a single jubilant conference experience, followed by a lifetime of being shunned at all remaining professional meetings. If we could infer the population parameters, given the sample statistics and sample size, then we might have some confidence that future research would yield sample statistics similar to those in our own sample.

Unfortunately, the direction of the inference in inferential statistics is *from* the population and *to* the sample, and *not* from the sample to the population (Thompson, 1997). Thus Cohen (1994) concluded that the statistical significance test 'does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!' (p. 997).

Recommended Changes in Practice

A few scholars have called for the banning of statistical significance tests (cf. Carver, 1978, 1993). However, the fact that many psychologists misinterpret statistical significance tests is not a reasonable warrant for banning these tests. As Strike (1979) explained, 'To deduce a proposition with an "ought" in it from premises containing only "is" assertions is to get something in the conclusion not contained in the premises, something impossible in a valid deductive argument' (p. 13). In logic this fallacy is called a 'should/would' or 'is/ought' error (Hudson, 1969).

But more and more researchers also now realize that 'virtually any study can be made to show [statistically] significant results if one uses enough subjects' (Hays, 1981, p. 293). This means that

> Statistical significance testing can involve a tautological logic in which tired researchers, having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they're tired. (Thompson, 1992b, p. 436)

Consequently, attention has now turned toward ways to improve practice. Five potential improvements in practice are suggested here.

Effect Size Reporting

Empirical studies of articles published since 1994 in psychology, counseling, special education and general education suggest that merely '*encouraging*' effect size reporting (APA, 1994) has *not* appreciably affected actual reporting practices (e.g Kirk, 1996; Snyder & Thompson, 1998; Thompson & Snyder, 1997, 1998; Vacha-Haase & Nilsson, 1998). Apparently, when it comes to reporting and interpreting effect sizes, many are called but few choose to be chosen. Consequently, editorial policies at some journals now *require* authors to report and interpret effect sizes (Heldref Foundation, 1997; Thompson, 1994b; see also Loftus, 1993; Shrout, 1997). It is particularly noteworthy that editorial policies even at one APA journal now indicate that

> If an author decides not to present an effect size estimate along with the outcome of a significance test, I will ask the author to provide specific justification for why effect sizes are not reported. So far, I have not heard

> a good argument against presenting effect sizes. Therefore, unless there is a real impediment to doing so, you should routinely include effect size information in the papers you submit. (Murphy, 1997, p. 4)

Effect sizes are important to report and interpret for at least two reasons. First, these indices can help inform judgment regarding the practical or substantive significance of results. Statistical significance tests do not bear upon the noteworthiness of results, because improbable events are not necessarily important (see Shaver's [1985] classic example), and because 'if the null hypothesis is not rejected, it is usually [only] because the *N* is too small' (Nunnally, 1960, p. 643).

Second, reporting effect sizes facilitates the meta-analytic integration of findings across a given literature. People who incorrectly believe, either consciously or unconsciously, that statistical significance tests evaluate the probability of population parameters can exaggerate the importance of a single study, because the study then generalizes to the population. Persons who recognize the limits of these statistical tests realize that most single studies are important primarily only as building blocks within a cumulative body of evidence. As Schmidt (1996) noted:

> Meta-analysis . . . has revealed how little information there typically is in any single study. It has shown that, contrary to widespread belief, a single primary study can rarely resolve an issue or answer a question. (p. 127)

Reporting effect sizes helps meta-analysts more easily and more accurately synthesize findings, because the analyst can then avoid using more approximate effects computed based on sometimes tenuous statistical assumptions.

Of course, effect size is no more a panacea than is a statistical significance test, for two reasons noted by Zwick (1997). First, because human values are also not part of the calculation of an effect size, any more than values are part of the calculation of *p*, 'largeness of effect does not guarantee practical importance any more than statistical significance does' (p. 4).

Second, some researchers seem to have adopted Cohen's (1988) definitions of small, medium and large effects with the same rigidity that '$\alpha = .05$' has been adopted. Such rigidity is inappropriate. Cohen only intended these as impressionistic characterizations of result typicality across a diverse literature, and not as rigid universal criteria. However, some empirical studies suggest that the characterization is reasonably accurate (Glass, 1979; Olejnik, 1984), at least as regards a literature historically built with a bias against statistically non-significant results (Rosenthal, 1979).

Notwithstanding these caveats, it is suggested that *all authors of quantitative studies should report and interpret effect sizes*. Because merely encouraging these practices has to date had little or no effect, at some point it may become necessary to require that effect sizes are reported. Of course, a requirement that effect sizes be reported does not inherently require that a whole new system of statistical analyses be invoked; all our classical analytic methods can be used to yield both $p_{\text{CALCULATED}}$ and effect size values, even though the methods have traditionally been used only for the first purpose.

Effect Size Interpretability

There are myriad effect sizes from which the researcher can choose. Useful reviews of the choices have been provided by Kirk (1996), Snyder and Lawson (1993) and Friedman (1968), among others.

Effect sizes can be categorized into two broad classes: variance-accounted-for measures (e.g. R^2, eta$(\eta)^2$) and standardized differences (e.g. Cohen's *d*, Hedges' *g*) (Kirk [1996] identifies a third, 'miscellaneous' class. Variance-accounted-for indices can be computed in all classical statistical analyses because all analyses are correlational, even though some designs are experimental and some are not (Knapp, 1978; Thompson, in press).

Furthermore, effect sizes can be further subdivided as being either 'uncorrected' (e.g. R^2, eta^2) or 'corrected' (e.g. adjusted R^2, omega$(\omega)^2$). Because all conventional analyses are least-squares correlational methods that capitalize on all sample variance, including the sampling error variance unique to the sample, all uncorrected variance-accounted-for statistics are positively biased and overestimate population effects. This bias can be statistically removed via the corrected effect size formulas which estimate the influence of the three major factors contributing to sampling error:

1. Samples with smaller sample sizes tend to have more sampling error.
2. Studies with more variables tend to have more sampling error.
3. Samples from populations with larger variance-accounted-for parameters tend to have less sampling error.

Regarding this last influence, the case can be made clear at the extreme for a study involving the statistic r^2. If the population parameter is 1.0, it is impossible to draw a sample that yields an inaccurate effect size, since from this population every sample involving any number of pairs of scores will yield an r^2 of 1.0.

The field has not yet established a single preferred effect size, a preference for variance-accounted-for as against standardized differences indices, or a preference for corrected as against uncorrected indices. It is doubtful that the field will ever settle on a single index to be used in all studies, given that so many choices exist and because the statistics can usually be translated into approximations across the two major classes. However, some pluses and minuses for both variance-accounted-for and standardized differences indices can be noted.

On the one hand, variance-accounted-for indices do have the benefit of reinforcing the realization that all classical analyses are correlational

(Knapp, 1978; Thompson, in press). This may minimize the autonomic choice of ANOVA as an analytic method based on an unconscious association of ANOVA with the ability to make causal inferences (cf. Humphreys & Fleishman, 1974).

On the other hand, standardized difference effect sizes (e.g. the difference of the experimental group mean minus the control group mean divided by the control group standard deviation) may be more directly interpretable. For example, Saunders, Howard and Newman (1988) argued that a variance-accounted-for effect is 'still cast in a language that was foreign to (and unusable by) practitioners' (pp. 207–208); a variance-accounted-for 2 percent effect usually must be expressed in the metric of an outcome variable to be meaningful.

However, not all studies involve experiments or a focus on means, and the use of standardized differences can seem stilted in such contexts. Thus, there are no clear-cut choices of an optimal effect size, or even a class of effect indices.

But it does seem reasonable to expect at a minimum that *effect sizes should always be presented in an accessible metric* (e.g. years added to longevity, on the average, from not smoking; median number of additional months due to an intervention that Alzheimer's patients were able to live without institutionalization). Several clinical disciplines have explored innovative ways to meet these requirements (see, e.g., the half-dozen articles in a 1988 special issue of *Behavioral Assessment*, including the report by Saunders et al. [1988]). But continued development of more effective ways to communicate effects remains warranted.

Values Explication

Cohen's (1988) typicality characterizations are not suitable as rigid criteria for noteworthiness, nor were they meant to be so used. The only suitable criteria for evaluating result value (a) must be informed by the personal, idiosyncratic values of each researcher and (b) must take into account the particular context of a given study. Regarding the first point, Huberty and Morris (1988) noted that, 'As in all of statistical inference, subjective judgment cannot be avoided. Neither can reasonableness!' (p. 573).

Regarding the context of a given study, a 2 percent variance-accounted-for effect size will not be noteworthy to most researchers (or to most readers) in the context of a study like one I once read titled 'Smiling and Touching Behavior of Adolescents in Fast Food Restaurants'. However, Gage (1978) pointed out that the relationship between cigarette smoking and lung cancer involves roughly this same effect size, and noted that:

> Sometimes even very weak relationships can be important. . . . [O]n the basis of such correlations, important public health policy has been made and millions of people have changed strong habits. (p. 21)

Certainly a small variance-accounted-for effect size involving highly valued outcomes, such as longevity, can be noteworthy. But since the judgments of result noteworthiness are inherently value-driven, and are 'on the average', even here some may reach a seemingly reasoned decision that the effect is not noteworthy, or at least not noteworthy enough to merit changed behavior.

Many scientists will probably feel uncomfortable declaring their effects in a meaningful metric and then explicating the associated personal or societal values that make these effects noteworthy. Declarations that 'My results were [statistically] significant' will have to be replaced with, 'This intervention extends life expectancy, on the average, by 1.4 years, and given my valuing of life, I believe this result is noteworthy.'

Historically, social scientists have used *p* statistics as a way to finesse values differences, because conflicting values of different people are not readily reconcilable. Nevertheless, *researchers should be expected to declare the values that make their effects noteworthy*.

Normative practices for evaluating such assertions will have to evolve. Research results should not be published merely because the individual researcher thinks the results are noteworthy. By the same token, editors should not quash research reports merely because they find explicated values unappealing. These resolutions will have to be formulated in a spirit of reasoned comity.

But we also must realize that our historical reliance on *p* values as a way to avoid value assertions led only to feigned objectivity, and not to real objectivity. This feigned objectivity was built on the edifice of misinterpretation of what statistical significance tests really do.

Evidence of Replicability

The cumulation of knowledge about relationships that recur under specified conditions is the sine qua non of science for those psychologists who believe that such laws can reasonably be formulated. For these psychologists evidence of result replicability is critical for creating a warrant that results are noteworthy.

The required nature of this warrant has received too little attention in an era when statistical significance tests were thought to evaluate result replicability, when these tests were thought to evaluate (rather than merely to presume) selected population parameters. Several vehicles for establishing these warrants can be noted.

One warrant involves an important contribution that Jacob Cohen made in his 1994 article; this very important contribution has not been as widely noticed as might be hoped (Hagen, 1997). Cohen (1994) carefully distinguished the general class of 'null' hypothesis tests from a subclass of null tests he labeled the 'nil' hypothesis test. (A related important distinction is what Meehl [1997] has described as ‘strong’ vs ‘weak’ null hypothesis refutation.)

For Cohen, a nil null hypothesis always specifies zero difference or zero relationship (e.g. for the especially inappropriate test of a reliability statistic, H_0: $r_{XX} = 0$; H_A: $r_{XX} \neq 0$), while other non-nil null hypotheses may test an alternative hypothesis such as H_A: $r_{XX} > .7$). Cohen’s important distinction recognizes that a ‘null hypothesis means the hypothesis to be nullified, not necessarily a hypothesis of no difference’ (Chow, 1988, p. 105).

Some specific null must be presumed true in the population, or otherwise infinitely many parameters are possible and the $p_{\text{CALCULATED}}$ for the sample results becomes indeterminate (Thompson, 1996). Most researchers use a nil hypothesis as the null partly because this is what most computer packages assume, and partly because methodology for invoking non-nil null hypotheses has some ‘complexity, and it is not yet readily applicable in many designs’ (Dar, Serlin, & Omer, 1994, p. 81).

The mindless use of the nil hypothesis obviates the necessity prospectively to extrapolate thoughtful expected effect sizes from prior literature as part of study design. Furthermore, the interpretation of ‘[statistical] significance’ as indicating result value means that some researchers do not retrospectively interpret their study effects in the context of specific previous findings. These failures are most unfortunate, because the prospective and retrospective use of effects from prior studies is itself a check on the replicability of results in a given inquiry.

Empirical evidence for result replicability can be either ‘external’ or ‘internal’ (Thompson, 1993, 1996). ‘External’ replication studies invoke a new sample measured at a different time and/or a different location. Such replications have unfortunately been undervalued (Robinson & Levin, 1997), perhaps because some researchers thought they were already testing replicability by conducting statistical significance tests.

‘Internal’ replicability analyses use the sample in hand to combine the participants in different ways to try to estimate how much the idiosyncracies of individuality within the sample have compromised sample results. The major ‘internal’ replicability analyses are cross-validation, the jackknife and the bootstrap (Diaconis & Efron, 1983); the logics are reviewed in more detail elsewhere (cf. Thompson, 1993, 1994c).

‘Internal’ evidence for replicability is never as good as an actual replication (Robinson & Levin, 1997; Thompson, 1997), but is certainly better than presuming that a statistical significance test assures result replicability. And such ‘internal’ replicability evidence is useful for researchers who for practical reasons cannot externally replicate all results prior to graduation or tenure review.

It is important that, when used to evaluate result replicability, these logics are not confused with other uses of the same logics (Thompson, 1993). For example, the *inferential* use of the bootstrap involves using the bootstrap to estimate a sampling distribution when the sampling distribution is not known or assumptions for the use of a known sampling distribution cannot be met. The *descriptive* use of the bootstrap looks primarily at the variance in parameter estimates across many different combinations of the participants.

The inferential application requires considerably more ‘re-samples’ (see Thompson, 1994c) than the descriptive application recommended here. This is because the inferential focus is on the tails of the estimated sampling distribution (e.g. the 95th percentile of the distribution, for a one-tailed statistical significance test), rather than the descriptive focus on the standard deviation (i.e. the ‘standard error’) of the sampling distribution. Participants in the tails of the sampling distribution are rarer, and therefore many more bootstrap re-samples are required to estimate these very small or large percentiles.

The field has not yet resolved all the issues involved in establishing a sufficient warrant for result replicability, again, perhaps, because some authors incorrectly assumed that statistical tests evaluated the population. The relevant software to conduct ‘internal’ bootstrap analyses is already available (e.g. Lunneborg, 1987, for univariate applications, and Thompson, 1992a, 1995, for multivariate applications). Because replicability evidence is critical to the cumulation of knowledge, *more authors should be expected to provide some evidence of result replicability*.

Reporting Confidence Intervals

Various scholars have recommended that confidence intervals should be used to replace or supplement statistical significance tests (e.g. Dar, Serlin, & Omer, 1994; Meehl, 1997; Schmidt, 1996; Serlin, 1993). However, researchers using confidence intervals must remember that ‘the interval endpoints are themselves random variables’ (Zwick, 1997, p. 5) also estimated using sample data. That is, the confidence interval does *not* indicate that, given the endpoints, the chances are *X* percent that the interval will include the parameter (Falk & Greenbaum, 1995; Howson & Urbach, 1994). Furthermore, researchers who mindlessly interpret confidence intervals only against the standard of whether the interval subsumes zero are doing nothing more than a mindless ‘nil’ hypothesis test (Cortina & Dunlap, 1997).

However, confidence intervals do have one very appealing feature, as Schmidt (1996) made clear. Even if all the research in an area of inquiry was based on radically erroneous estimates of parameters (and even if these a priori estimates were used in specifying non-nil null hypotheses), the parameter would still emerge across studies as a series of overlapping confidence intervals converging on the same parameter.

The use of confidence intervals might also mitigate against the current bias in the literature (a) first favoring the publication of Type I errors and (b) then disfavoring publication of replication studies revealing the previously published Type I error. Setting alpha at a small level does not prevent any Type I errors; rather, the percentage of such errors is capped at a small proportion. But some such errors will unavoidably occur. Because the literature has been biased in favor of statistically significant results (Rosenthal, 1979), such Type I errors are afforded priority for publication, but the replications with statistically non-significant results will compete at a disadvantage for journal space, and so the self-correction of science through replication will be impeded. Greenwald (1975) cited relevant actual examples.

A focus on consistency of findings across studies can be achieved with confidence intervals interpreted in relation to each other, rather than against the nil standard of a zero value. Therefore, it is suggested that *more authors should report confidence intervals as part of their results*.

Summary

Kirk (1996) recently noted that, ‘Our science has paid a high price for its ritualistic adherence to null hypothesis significance testing’ (p. 756). The overuse and misinterpretation of statistical tests has been frequently decried as well in literatures other than psychology, including medicine (Kraemer, 1992; Pocock, Hughes, & Lee, 1987), business (Sawyer & Peter, 1983), occupational therapy (Ottenbacher, 1984) and speech and hearing (Young, 1993). Nevertheless, the use of statistical significance tests remains common, and some empirical studies reflect even an increased use of these methods (Parker, 1990)!

Many have marveled at the robustness of the statistical significance logic against the application of the wooden stake through the heart. For example, Falk and Greenbaum (1995) noted:

> We have shown the compelling nature and the robustness of that illusion [that statistical significance tests give us the information we need]. A massive educational effort is required to eradicate the misconception and extinguish the mindless use of a procedure that dies hard. (p. 94)

And Harris (1991) observed, ‘it is surprising that the dragon will not stay dead’ (p. 375).

Frick (1996) cited an anonymous reviewer of his defense of statistical significance testing who argued that, ‘A way of thinking that has survived decades of ferocious attacks is likely to have some value’ (p. 379). Of course, this view presumes a completely rational model of science in which scientists are objective, dispassionate logicians never acting merely out of habit; the view also presumes that scientists are always anxious to admit past errors publicly made in the articles they themselves published over the courses of their careers.

Five specific suggestions for improved analytic practice have been presented here. It should be noted that these suggestions can be followed even by those psychologists still employing conventional statistical significance tests. But social science will proceed most rapidly when research becomes the search for replicable effects noteworthy in magnitude in the context of both the inquiry and personal or social values.

Note

1. For each of the three pairs of studies, the first study within each pair has a smaller $p_{\text{CALCULATED}}$ value, if conventional nil null hypotheses (i.e. H_0: $M_1 = M_2 = M_3$; H_0: $SD_1 = SD_2$; and $R^2 = 0$) are used.

References

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

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

Carver, R. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, *61*, 287–292.

Chow, S.L. (1988). Significance test or effect size? *Psychological Bulletin*, *103*, 105–110.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312.

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

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

Cronbach, L.J. (1975). Beyond the two disciplines of psychology. *American Psychologist*, *30*, 116–127.

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist*, *42*, 145–151.

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.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, *248*(5), 116–130.

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

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

Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin*, *70*, 245–251.

Gage, N.L. (1978). *The scientific basis of the art of teaching*. New York: Teachers College Press.

Glass, G.V. (1979). Policy for the unpredictable (uncertainty research and policy). *Educational Researcher*, *8*(9), 12–14.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Greenwald, A.G., Gonzalez, R., Harris, R.J., & Guthrie, D. (1996). Effect size and *p*-values: What should be reported and what should be replicated? *Psychophysiology*, *33*, 175–183.

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

Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Harris, M.J. (1991). Significance tests are not enough: The role of effect size estimation in theory corroboration. *Theory & Psychology*, *1*, 375–382.

Hays, W.L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart & Winston.

Heldref Foundation. (1997). Guidelines for contributors. *Journal of Experimental Education*, *65*, 287–288.

Howson, C., & Urbach, P. (1994). Probability, uncertainty and the practice of statistics. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 39–51). Chichester: Wiley.

Huberty, C.J., & Morris, J.D. (1988). A single contrast test procedure. *Educational and Psychological Measurement*, *48*, 567–578.

Hudson, W.D. (1969). *The is/ought question*. London: Macmillan.

Humphreys, L.G., & Fleishman, A. (1974). Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. *Journal of Educational Psychology*, *66*, 464–472.

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

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–759.

Knapp, T.R. (1978). Canonical correlation analysis: A general parametric significance testing system. *Psychological Bulletin*, *85*, 410–416.

Kraemer, H.C. (1992). Reporting the size of effects in research studies to facilitate assessment of practical or clinical significance. *Psychoendocrinology*, *17*, 527–536.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. *American Psychologist*, *43*, 635–642.

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

Lunneborg, C.E. (1987). *Bootstrap applications for the behavioral sciences*. Seattle: University of Washington Press.

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

Meehl, P.E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 391–423). Mahwah, NJ: Erlbaum.

Morrison, D.E., & Henkel, R.E. (Eds.). (1970). *The significance test controversy*. Chicago, IL: Aldine.

Murphy, K.R. (1997). Editorial. *Journal of Applied Psychology, 82*, 3–5.

Nelson, N., Rosenthal, R., & Rosnow, R.L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, *41*, 1299–1301.

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 behavioral sciences*. New York: Wiley.

Olejnik, S.F. (1984). Planning educational research: Determining the necessary sample size. *Journal of Experimental Education*, *53*, 40–48.

Ottenbacher, K. (1984). Measures of relationship strength in occupational therapy research. *The Occupational Therapy Journal of Research*, *4*, 271–285.

Parker, S. (1990). A note on the growth of the use of statistical tests in *Perception & Psychophysics. Bulletin of the Psychonomic Society, 28*, 565–566.

Pedhazur, E.J., & Schmelkin, L.P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

Pocock, S.J., Hughes, M.D., & Lee, R.J. (1987). Statistical problems in the reporting of clinical trials. *The New England Journal of Medicine*, *317*, 426–432.

Robinson, D., & Levin, J. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, *26*(5), 21–26.

Rosenthal, R. (1979). The ‘file drawer problem’ and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rosenthal, R. (1991). Effect sizes: Pearson’s correlation, its display via the BESD, and alternative indices. *American Psychologist*, *46*, 1086–1087.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, *55*, 33–38.

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

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

Saunders, S.M., Howard, K.I., & Newman, F.L. (1988). Evaluating the clinical significance of treatment effects: Norms and normality. *Behavioral Assessment*, *10*, 207–218.

Sawyer, A.G., & Peter, J.P. (1983). The significance of statistical significance tests in marketing research. *Journal of Marketing Research*, *20*, 122–123.

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

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

Serlin, R.C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *Journal of Experimental Education*, *61*, 350–360.

Shaver, J. (1985). Chance and nonsense. *Phi Delta Kappan, 67*, 57–60.

Shea, C. (1996). Psychologists debate accuracy of ‘significance test’. *Chronicle of Higher Education, 42,* A12, A16.

Shrout, P.E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science, 8,* 1–2.

Snyder, P.A., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61,* 334–349.

Snyder, P.A., & Thompson, B. (1998). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested alternatives. *School Psychology Quarterly, 13,* 335–348.

Strike, K.A. (1979). An epistemology of practical research. *Educational Researcher, 8*(1), 10–16.

Thompson, B. (1992a). DISCSTRA: A computer program that computes bootstrap resampling estimates of descriptive discriminant analysis function and structure coefficients and group centroids. *Educational and Psychological Measurement, 52,* 905–911.

Thompson, B. (1992b). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70,* 434–438.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61,* 361–377.

Thompson, B. (1994a). The concept of statistical significance testing (An ERIC/AE Clearinghouse Digest #EDO-TM-94-1). *Measurement Update, 4,* 5–6. (ERIC Document Reproduction Service No. ED 366 654)

Thompson, B. (1994b). Guidelines for authors. *Educational and Psychological Measurement, 54,* 837–847.

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

Thompson, B. (1995). Exploring the replicability of a study’s results: Bootstrap statistics for the multivariate case. *Educational and Psychological Measurement, 55,* 84–94.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26–30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26*(5), 29–32.

Thompson, B. (1998a). Review of *What if there were no significance tests? Educational and Psychological Measurement, 58.*

Thompson, B. (1998b, January). *Why ‘encouraging’ effect size reporting isn’t working: The etiology of researcher resistance to changing practices.* Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX. (ERIC Document Reproduction Service No. ED 416 214)

Thompson, B. (in press). Canonical correlation analysis. In L. Grimm & P. Yarnold (Eds.), *Reading and understanding multivariate statistics, Vol. 2.* Washington, DC: American Psychological Association.

Thompson, B., & Snyder, P.A. (1997). Statistical significance testing practices in the *Journal of Experimental Education. Journal of Experimental Education, 66,* 75–83.

Thompson, B., & Snyder, P.A. (1998). Statistical significance and reliability analyses in recent *JCD* research articles. *Journal of Counseling and Development, 76,* 436–441.

Vacha-Haase, T., & Nilsson, J.E. (1998). Statistical significance reporting: Current trends and usages within *MECD*. *Measurement and Evaluation in Counseling and Development*, *31*, 46–57.

Young, M.A. (1993). Supplementing tests of statistical significance: Variation accounted for. *Journal of Speech and Hearing Research*, *36*, 644–656.

Zuckerman, M., Hodgins, H.S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science*, *4*, 49–53.

Zwick, R. (1997, March). *Would the abolition of significance testing lead to better science?* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

ACKNOWLEDGEMENTS. I appreciate the thoughtful comments of Mark Applebaum, Jacob Cohen, Lee Cronbach, Ruma Falk, Carl Huberty, Joel Levin, Paul Meehl, Frederick Mosteller, Frank Schmidt and Rebecca Zwick on a previous draft of this paper.

BRUCE THOMPSON is Professor and Distinguished Research Scholar, Department of Educational Psychology, Texas A&M University, and Adjunct Professor of Community Medicine, Baylor College of Medicine (Houston, TX). He is a member of the APA Task Force on Statistical Inference. He is the author of 133 articles, and author/editor of six books. He previously edited *Measurement and Evaluation in Counseling and Development*, and currently edits *Educational and Psychological Measurement*, the *Journal of Experimental Education* and the JAI Press book series ‘Advances in Social Science Methodology’. ADDRESS: Department of Educational Psychology, Texas A&M University, College Station, TX 77843–4225, USA. The author and related reprints can both be accessed on the Internet via URL: http://acs.tamu.edu/~bbt6147/