DOCUMENT RESUME

ED 392 819 TM 024 458

AUTHOR Thompson, Bruce
TITLE Editorial Policies Regarding Statistical Significance Testing: Three Suggested Reforms.
PUB DATE 8 Nov 95
NOTE 24p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Biloxi, MS, November 1995).
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Editing; *Educational Assessment; *Effect Size; Quality Control; *Research Methodology; *Statistical Significance; *Test Use
IDENTIFIERS Bootstrap Methods; Cross Validation; Jackknifing Technique; *Research Replication

ABSTRACT

Editorial practices revolving around tests of statistical significance are explored. The logic of statistical significance testing is presented in an accessible manner--many people who use statistical tests might not place such a premium on them if they knew what the tests really do, and what they do not do. The etiology of decades of misuse of statistical tests is explored, highlighting the bad implicit logic of persons who misuse statistical tests. Finally, three revised editorial policies that would improve conventional practice are discussed. The first is the use of better language, with insistence on universal use of the phrase "statistical significance" to emphasize that the common meaning of "significant" has nothing to do with results being important. A second improvement would be emphasizing effect size interpretation, and a third would be using and reporting strategies that evaluate the replicability of results. Internal replicability analyses such as cross validation, the jackknife, or the bootstrap would help determine whether results are stable across sample variations. (Contains 51 references.) (Author/SLD)

aerasign.rev 9/13/95

Editorial Policies Regarding Statistical Significance Testing: Three Suggested Reforms

Bruce Thompson

Texas A&M University
and
Baylor College of Medicine

RUNNING HEAD: Editorial Policies

Author Address:
Bruce Thompson
TAMU Dept Educ Psyc
College Station, TX 77843-4225
Voice 409/845-1335
FAX 409/862-1256
e-mail: E100BT@TAMVM1.TAMU.EDU

Paper presented at the annual meeting of the Mid-South Educational Research Association, Biloxi, MS, November 8, 1995.

Abstract

The present comment reviews practices revolving around tests of statistical significance. First, the logic of statistical significance testing is presented in an accessible manner; many people who use statistical tests might not place such a premium on the tests if these individuals understood what the tests really do, and what the tests do not do. Second, the etiology of decades of misuse of statistical tests is briefly explored; we must understand the bad implicit logic of persons who misuse statistical tests if we are to have any hope of persuading them to alter their practices. Third, three revised editorial policies that would improve conventional practice are highlighted.

The recently published American Psychological Association (APA) style manual includes an important, but largely unheralded, shift in APA editorial policy regarding the use of statistical significance testing in quantitative research. The manual notes that:

> Neither of the two types of probability values reflects the importance or magnitude of an effect because both depend on sample size... You are encouraged to provide effect-size information. (APA, 1994, p. 18)

This shift in emphasis contrasts sharply with traditional editorial practice within many journals in the behavioral sciences. For example, after 12 years as editor of the Journal of Experimental Psychology, Melton boasted that:

> In editing the Journal there has been a strong reluctance to accept and publish results related to the principal concern of the researcher when those results were [statistically] significant [only] at the .05 level... It reflects a belief that it is the responsibility of the investigator in a science to reveal his [sic] effect in such a way that no reasonable man [sic] would be in a position to discredit the results by saying that they were the product of the way the ball bounces. (Melton, 1962, p. 554)

The shift of emphasis toward effect size and replicability analysis, at the expense of emphasis on statistical significance testing, certainly did not occur overnight. APA's flagship journal, the American Psychologist, first included a seemingly periodic series of articles on the extraordinary limits of statistical significance testing (cf. J. Cohen, 1990, 1994; Kupfersmid, 1988; Rosenthal, 1991; Rosnow & Rosenthal, 1989).

Of course, these views are hardly new. A few especially noteworthy among the numerous efforts "to exorcise the null hypothesis" (Cronbach, 1975, p. 124) over the past 35 years have been works by Rozeboom (1960), Morrison and Henkel (1970), Carver (1978), Meehl (1978), Shaver (1985), Oakes (1986), and J. Cohen (1994). The entire Volume 61, Number 4 issue of the Journal of Experimental Education was devoted to these themes.

However, a perusal of AERA publications and of papers presented at our annual meetings does not suggest that old knowledge in this area has yet had major impacts on contemporary practice. The message apparently has not yet been clearly put in AERA forums, or in any case seems to bear reiteration.

The present brief essay has three purposes. First, the logic of statistical significance testing is presented in an accessible manner; many people who use statistical tests might not place such a premium on the tests if these individuals understood what the tests really do, and what the tests do not do. Second, the etiology of decades of misuse of statistical tests is briefly explored; we must understand the bad implicit logic of persons who misuse statistical tests if we are to have any hope of persuading them to alter their practices--it will not be sufficient merely to tell researchers not to use statistical tests, or to use them more judiciously. Third, revised editorial policies that would focus interpretations on noteworthy results (i.e., findings not involving statistical significance testing) are highlighted.

The Logic of Statistical Testing

The use of statistical significance testing logic dates back almost 300 years to studies of birth rates by John Arbuthnot in 1710 (Huberty, 1993). However, use of variations on these tests were popularized in the social sciences by Sir Ronald Fisher and by Jerzy Neyman and Egon Pearson (Huberty, 1987). Today, most researchers implicitly employ some hybrid of the logics suggested by these three figures, but the logics all involve the computation of some form of $p_{CALCULATED}$.

Because p values are difficult to compute, researchers traditionally have conducted statistical tests by invoking test statistics, such as F or t. Using test statistics always yields that same decisions as does the use of p values, but test statistics are easier to compute. Of course, today these computational advantages of test statistics have now been rendered moot by modern computer software that routinely provides exact $p_{CALCULATED}$ values, and so researchers are no longer yoked to the use of the conventional p values (e.g., .05 and .01) for which the related test statistic critical values are widely published.

Unfortunately, very few researchers seem to understand what their p calculated values actually evaluate (Carver, 1978). Put succinctly, $p_{CALCULATED}$ is *the probability (0 to 1.0) of the sample statistics, given the sample size, and assuming the sample was derived from a population in which the null hypothesis (H_0) is exactly true* (Thompson, 1994a). The computation of $p_{CALCULATED}$ in a particular study includes consideration of three elements: (a) the results in the sample (i.e., the *sample "statistics"*) vis a vis the null hypothesis (i.e., sample means, medians, standard deviations, or whatever a given null hypothesis is about); (b) the related results in the population (i.e., the *population "parameters"*) vis a vis the null hypothesis (i.e., population means, medians, standard deviations, or whatever a given null hypothesis is about); and (c) the *sample size*.

For example, let's presume a researcher has a sample of scores on a reading ability test ($\underline{X}$) for two groups of subjects, and wants to test whether the "spreadoutness" of the scores in the two groups is equal. Perhaps in group one $\underline{SD}_X$ is 3, and in group two $\underline{SD}_X$ is 5. The researcher wants to know the probability of obtaining standard deviations of 3 and 5 (these sample standard deviations of 3 and 5 are called "statistics"), respectively, assuming the samples came from a population in which the two standard deviations (these population standard deviations are called "parameters") were equal.

Why must the researcher assume that the sample comes from a population in which H_0 is true? Well, something must be assumed, or there would be infinitely many equally plausible (i.e., indeterminate) answers to the question of what is the probability of the sample statistics. For example, sample statistics of standard deviations of 3 and 5 would be most likely (highest $p_{CALCULATED}$) if the population parameter standard deviations were 3 and 5, would be slightly less likely if the population standard deviations were 3.3 and 4.7, and would be less likely still (an even smaller $p_{CALCULATED}$) if the parameters were standard deviations of 4 and 4.

Researchers can assume that any population parameters, as long as they make some specific assumptions regarding what the parameters are. However, almost all statistical packages (and consequently almost all researchers) assume that an H_0 of "no difference" is true in the population.

But why must computations of $p_{CALCULATED}$ take into account the researcher's sample size? The answer is that sample statistics other than those that exactly honor the null hypothesis are less and less likely (i.e., yield smaller and smaller $p_{CALCULATED}$ values) as the sample size increases. For example, sample standard deviations of 3 and 5 really could come from a population with standard deviation parameters of 4 and 4. But such a possibility is more likely if sample size is small, because smaller sample sizes have more "sampling error" or "flukiness" in them. Therefore, since a sample deviation from equality would be more likely with a small sample of six people in each group, the $p_{CALCULATED}$ for these statistics for this sample size would be larger. But as sample size got larger for the same statistics (e.g., sample standard deviations of 3 and 5), the $p_{CALCULATED}$ values would get smaller and smaller.

One serious problem with this statistical testing logic is that the in reality H_0 is never true in the population, as recognized by any number of prominent statisticians (Tukey, 1991), i.e., there will always be some differences in population parameters, although the differences may be incredibly trivial. Near 40 years ago Savage (1957, pp. 332-333) noted that, "Null hypotheses of no difference are usually known to be false before the data are collected." Subsequently, Meehl (1978, p. 822) argued, "As I believe is generally recognized by statisticians today and by thoughtful social scientists, the null hypothesis, taken literally, is always false." Similarly, noted statistician Hays (1981, p. 293) pointed out that "[t]here is surely nothing on earth that is completely independent of anything else. The strength of association may approach zero, but it should seldom or never be exactly zero." And Loftus and Loftus (1982, pp. 498-499) argued that, "finding a '[statistically] significant effect' really provides very little information, because it's almost certain that some relationship (however small) exists between any two variables."

The very important implication of all this is that statistical significance testing primarily becomes only a test of researcher endurance, because "virtually any study can be made to show [statistically] significant results if one uses enough subjects" (Hays, 1981, p. 293). As Nunnally (1960, p. 643) noted some 35 years ago, "If the null hypothesis is not rejected, it is usually because the N is too small. If enough data are gathered, the hypothesis will generally be rejected." The implication is 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. This tautology has created considerable damage as regards the cumulation of knowledge... (Thompson, 1992, p. 436)

The Etiology of Statistical Testing

The etiology of the propensity to conduct statistical significance tests can be traced to two dynamics. The first involves an unrecognized error in logic when consciously trying to be scientific, while the second dynamic occurs as a frankly irrational process. These two dynamics undergirding continued emphasis on statistical tests must be understood if reform efforts are to be effective.

p as a Test of Result Replicability

The behaviors of many researchers, even some who protest otherwise, suggest erroneous beliefs (Shaver, 1993) that smaller $p_{CALCULATED}$ values mean that increasingly greater confidence can be vested in a conclusion that sample results are replicable. These researchers invoke a usually subliminal syllogism that takes the following form:

1. Small $p_{CALCULATED}$ means that ("A") sample statistics are at least approximately the ("B") population parameters (major premise);
2. The ("C") statistics for future samples drawn from the same population will approximate the ("B") population parameters (minor premise); so therefore,
3. The initial ("A") sample statistics will be replicated in the form of the ("C") statistics for future samples drawn from the same population (conclusion).

Their is no error in the deductive logic itself yielding the conclusion in this syllogism, since if "A"="B", and if "B"="C", then "A" does lead to "C". Nor is the minor premise of the syllogism incorrect.

But, as we have seen, statistical tests say "given an assumption about the parameters 'B', what is the likelihood of 'A', the sample statistics?", and not "given the sample statistics 'A', are these sample statistics likely 'B', the population parameters?". Carver (1978) cited myriad statistics textbooks that make precisely this logic error, and recent texts also illustrate related errors (Thompson, 1987, 1988). Carver (1978) argued that if our most respected scholars and teachers make this error so commonly, that therefore a *fortiori* there is less hope that the rest of us will avoid these pitfalls.

p as a Vehicle to Avoid Judgment

Too many researchers also believe that a statistically significant result is inherently important. These res archers erroneously equate an unlikely result with an inherently interesting result. Shaver's (1985, p. 58) classic example illustrates the folly of this equation in his hypothetical dialogue between two teachers:

> Chris: ...I set the level of significance at .05, as my advisor suggested. So a difference that large would occur by chance less than five times in a hundred if the groups weren't really different. An unlikely occurrence like that *surely* must be important.
>
> Jean: Wait a minute, Chris. Remember the other day when you went into the office to call home? Just as you completed dialing the number, your little boy picked up the phone to call someone. So you were connected and talking to one another without the phone ever ringing... Well, that must have been a truly important occurrence then?

Put simply, too many researchers wish to employ the mathematical calculation of probabilities only as a purely atavistic escape (a la Fromme's Escape from Freedom) from the existential human responsibility for making value judgments. But regrettably, as Daniel (1977, p. 425) noted,

> Whether or not the magnitude of the difference between *Mu* of A and *Mu* of B is of any practical importance is a question that cannot be answered by the statistical test. This is a question that only

> the researcher can answer after consideration of nonstatistical information.

Thompson (1993, p. 365) explained, "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."

Empirical science in inescapably a subjective business. As Berger and Berry (1988) argued, "objectivity is not generally possible in statistics" (p. 165). Huberty and Morris (1988, p. 573) concurred, noting that "As in all of statistical inference, subjective judgment cannot be avoided. Neither can reasonableness!"

Three Recommendations for Improved Editorial Policy

In evaluating statistical practices it is important to avoid making what in logic is termed an "is/ought" or a "should/would" error (Hudson, 1969; Hume, 1957). 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)

The fact that many researchers "are" now inappropriately using tests of statistical significance does not necessarily mean that researchers "ought" to abandon statistical tests.

However, various improvements in practice can certainly be recommended. For example, if researchers feel they must invoke statistical tests, then tests presuming null hypotheses of no difference might be eschewed in favor of tests postulating particular parameters based on previous research or on theory. Authors might also report "what if" analyses indicating at what different sample size a given fixed effect would become statistically significant, or would have no longer been statistically significant (cf. Thompson, 1989).

But the business of cumulating evidence about relationships that replicate under stated conditions would not be appreciably hindered by abandoning tests of statistical significance. Some acolytes argue that statistical tests are informative when findings are counter-intuitive (e.g., a statistically significant result is garnered with a small sample size), but the interpretation of effect sizes would equally well (and more directly) cue the researcher regarding the noteworthiness of such anomalous results.

Continued obsession with statistical significance would maintain current editorial practices favoring articles that report statistically significant outcomes (Rosenthal, 1979). The "file drawer" problem (Atkinson, Furlong & Wampold, 1982; L.H. Cohen, 1979; Greenwald, 1975) does create a fortunate bias against reports of Type II errors, since by definition statistically significant results cannot represent Type II errors. However, the bias toward statistically significant findings also creates a mentality where power is not reported (Olejnik, 1984) and is low (Woolley, 1983) in those few cases when results that are not statistically significant are published.

But, this bias also translates as a greater likelihood of reporting the rare statistically significant findings that are, in fact, actual Type I errors. Although researchers employ small alpha levels, some Type I errors will still be unavoidable across a large literature. This is problematic in the context of a bias against reporting results that are not statistically significant, "because investigators generally cannot get their failures to replicate published, [and so] Type I errors, once made, are very difficult to correct" (Clark, 1976, p. 258). Greenwald (1975, pp. 13-15) cites actual examples of such findings, the horrors of which Lindquist (1953, pp. 68-70) discussed some 40 years ago.

In any case, certain improvements in statistical routines should now be recognized as "best practice" by AERA editors, program chairs, and reviewers. At least three reforms should become explicit elements of AERA editorial practices.

Use of Better Language

If researchers are unable to report merely that they elected to reject a null hypothesis, such results ought to always be described as "statistically significant", and should never be described only as "significant." The universal use of the phrase, "statistically significant," might facilitate the recognition that the common meaning associated with "significant" has absolutely nothing to do with results being important (Carver, 1993), as explained previously.

Emphasizing Effect Size Interpretation

Several types of effect sizes can and should be reported and interpreted in all studies, regardless of whether statistical tests are or are not reported. AERA should venture beyond APA, and *require* such reports in all quantitative studies.

Classes of effect sizes include standardized differences (e.g., the experimental group mean minus the control group mean, divided by the estimated population standard deviation). Alternatively, since all analyses are correlational (cf. Knapp, 1978; Thompson, 1991), variance-accounted-for effect sizes can be computed in all studies. Either uncorrected effect sizes (e.g., R^2, eta^2) can be interpreted, or these can be corrected (e.g., omega^2, adjusted R^2) for the positive bias associated with (a) smaller sample sizes, (b) using more variables, and/or (c) smaller population effects. Snyder and Lawson (1993) present an understandable treatment of the choices.

Evaluating Result Replicability

If science is the business of discovering replicable effects, because statistical significance tests do not evaluate result replicability, then researchers should use and report some strategies that do evaluate the replicability of their results. Obviously, the only direct evaluation of result replicability is the so-called "external" replication (i.e., actual replication with a new sample). However, most researchers lack the stamina to conduct all their studies at least twice.

Researchers who find it difficult to replicate all their studies can use "internal" replicability analyses for this purpose. Such logics include using cross-validation, the jackknife, and/or the bootstrap. Thompson (1993, 1994b) provides an explanation of these empirical methods. Basically, the methods combine the subjects in hand in different ways to determine whether results are stable across sample variations, i.e., across the idiosyncracies of individuals which make generalization in social science so challenging.

Summary

For nearly 50 years, clarion calls for reformed practice regarding the use of statistical tests have been sounded. For example, some 45 years ago, prominent statistician Yates (1951, pp. 32-33) suggested that the use of statistical significance tests

> ...has caused scientific research workers to pay undue attention to the results of the tests of [statistical] significance they perform on their data, and too little to the estimates of the magnitude of the effects they are investigating... The emphasis on tests of [statistical] significance, and the consideration of the results of each experiment in isolation, have had the unfortunate consequence that scientific workers have often regarded the execution of a test of [statistical] significance on an experiment as the ultimate objective.

Bakan (1966, p. 436) noted almost 30 years ago, "When we reach a point where our statistical procedures are substitutes instead of aids to thought, and we are led to absurdities, then we must return to the common sense basis."

Meehl (1978, p. 817, 823) argued some 15 years ago:

> I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft [i.e., social science] areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology... I am not making some nit-picking statistician's correction. I am saying that the whole business is so radically defective as to be scientifically almost pointless.

And more recently, Dar (1987, p. 149) suggested that, "When passing null hypothesis tests becomes the criterion for successful predictions, as well as for journal publications, there is no pressure on the psychology researcher to build a solid, accurate theory; all he or she is required to do, it seems, is produce 'statistically significant' results."

Of course, editorial practices and policies have evolved somewhat, albeit incrementally. For example, the guidelines for authors of Measurement and Evaluation in Counseling and Development have for many years encouraged authors

> ...to assist readers in interpreting statistical significance of their results. For example, results may be indexed to sample size. An author may wish to say, "this correlation coefficient would have still been statistically significant even if sample size

> had been as small as *n* = 33," or "this correlation coefficient would have been statistically significant if sample size had been increased to *n* = 138." (Association for Assessment in Counseling, 1994, p. 143)

And the 1994 author guidelines for *Educational and Psychological Measurement* require authors to report and interpret effect sizes, and strongly encourage authors to report actual "external" replication studies, or to conduct "internal" replicability analyses.

The editorial practices within AERA would be improved if authors of articles and conference papers were encouraged (a) to correctly interpret statistical tests, (b) to always interpret effect sizes, and (c) to always explore result replicability. If our studies inform best practice in classrooms and other educational settings, the stakeholders in these locations certainly deserve better treatment from the research community via our analytic choices.

References

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

Association for Assessment in Counseling. (1994). Guidelines for authors. Measurement and Evaluation in Counseling and Development, 27(1), 341.

Atkinson, D.R., Furlong, M.J., & Wampold, B.E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? Journal of Counseling Psychology, 29, 189-194.

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

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

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(4), 287-292.

Clark, H.H. (1976). Reply to Wike and Church. Journal of Verbal Learning and Verbal Behavior, 15, 257-261.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312.

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

Cohen, L.H. (1979). Clinical psychologists' judgments of the scientific merit and clinical relevance of psychotherapy outcome research. Journal of Consulting and Clinical Psychology, 47, 421-423.

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

Daniel, W.W. (1977). Statistical significance versus practical significance. Science Education, 61, 423-427.

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

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

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

Huberty, C.J. (1987). On statistical testing. Educational Researcher, 16(8), 4-9.

Huberty, C.J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. Journal of Experimental Education, 61, 317-333.

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.

Hume, D. (1957). An inquiry concerning human understanding. New York: The Liberal Arts Press.

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

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

Lindquist, E.F. (1953). Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin.

Loftus, G.R., & Loftus, E.F. (1982). Essence of statistics. Monterey, CA: Brooks/Cole.

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

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

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

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.

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.

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.

Savage, R.J. (1957). Nonparametric significance. Journal of the American Statistical Association, 52, 331-344.

Shaver, J. (1985). Chance and nonsense. Phi Delta Kappan, 67(1), 57-60.

Shaver, J. (1993). What statistical significance testing is, and what it is not. Journal of Experimental Education, 61(4), 293-316.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61(4), 334-349.

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

Thompson, B. (1987). [Review of Foundations of behavioral research (3rd ed.)]. Educational Research and Measurement, 47, 1175-1181.

Thompson, B. (1988). [Review of Analyzing multivariate data]. Educational and Psychological Measurement, 48, 1129-1135.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and

Development, 22, 2-6.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling and Development, 24(2), 80-95.

Thompson, B. (1992). 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(4), 361-377.

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

Thompson, B. (1994b). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. Journal of Personality, 62(2), 157-176.

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

Woolley, T.W. (1983). A comprehensive power-analytic investigation of research in medical education. Journal of Medical Education, 85, 710-715.

Yates, F. (1951). The influence of Statistical methods for research workers on the development of the science of statistics. Journal of the American Statistical Association, 46, 19-34.