RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 33-38

Statistical Significance and Effect Size Reporting: Portrait of a Possible Future

Bruce Thompson
Texas A&M University and Baylor College of Medicine

The present paper comments on the matters raised regarding statistical significance tests by three sets of authors in this issue. These articles are placed within the context of contemporary literature. Next, additional empirical evidence is cited showing that the APA publication manual's "encouraging" effect size reporting has had no appreciable effect. Editorial policy will be required to affect change, and some model policies are quoted. Science will move forward to the extent that both effect size and replicability evidence of one or more sorts are finally seriously considered within our inquiry.

I appreciate the opportunity to comment on matters raised by Daniel (1998), McLean and Ernest (1998), and Nix and Barnette (1998) as regards statistical significance tests. Theme issues of journals such as the present one (see also Thompson (1993)) allow various perspectives to be articulated and help slowly but inexorably move the field toward improved practices. Of course, an important recent book (Harlow, Mulaik, & Steiger, 1997) also presents diverse perspectives regarding these continuing controversies (for reviews see Levin (1998) and Thompson (1998c)).

At the outset perhaps I should acknowledge possible conflicts of interest. First, co-editor Kaufman asked me to serve as one of the five or so referees who read each of these manuscripts in their initial form. Second, in a somewhat distant past, prior to his ascending to tenure, full professorship, and directorship of a research center, I chaired Larry Daniel's dissertation committee at the Uni-versity of New Orleans (boy, does reciting these facts make me feel old!).

These Articles and My Views in Context

It might be helpful to readers to frame these three articles, and my own views, within the context of views presented within the literature. Certainly at one extreme some authors (cf. Carver, 1978; Schmidt, 1996) have argued that statistical significance tests should be banned from publications. For example, Rozeboom (1997) recently argued that:

> Null-hypothesis significance testing is surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students . . . [I]t is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism . . . (p. 335)

Schmidt and Hunter (1997), virulent critics of statistical significance testing, similarly argued that, "Statistical significance testing retards the growth of scientific knowledge; it *never* makes a positive contribution" (p. 37, emphasis added).

At the other extreme (cf. Cortina & Dunlap, 1997; Frick, 1996), Abelson (1997) argued that, "Significance tests fill an important need in answering some key research questions, and if they did not exist they would have to be invented" (p. 118). Similarly, Harris (1997) argued that

> Null hypothesis significance testing (NHST) as applied by most researchers and journal editors can provide a very useful form of social control over researchers' understandable tendency to "read too much" into their data . . . [E]ven NHST alone would be an improvement over the current lack of attention to sampling error. (pp. 145, 164)

Bruce Thompson is a professor and distinguished research scholar in the Department of Educational Psychology at Texas A & M University. He is also an adjunct professor of community medicine at the Baylor College of Medicine. Correspondence regarding this article should be addressed to Bruce Thompson, Department of Educational Psychology, Texas A & M University, College Station, TX 77843-4225 or by e-mail to e100bt@tamvm1.tamu.edu. Related reprints and the author can be accessed on the Internet via URL: "**http://acs.tamu.edu/~bbt6147/**".

Some of these defenses of statistical tests have been thoughtful, but others have been flawed (Thompson, 1998b).

I see Nix and Barnette (1998) as somewhat approaching the Carver (1978)/Rozeboom (1997) end of the continuum. They "believe that most statisticians would [and seemingly should] welcome orderly change that would lead to abandonment of NHST." The authors feel constrained from supporting a ban, not on the merits, but only because of concerns regarding "democratic principles" and "censorship and infringement on individual freedoms."

McLean and Ernest (1998) believe that "our recommendations reflect a moderate mainstream approach." Certainly, their views are intellectually "moderate." A call that their views are "mainstream" requires a factual judgment as regards a moving target – our moving discipline. McLean and Ernest (1998) suggest that tests of statistical significance "must be accompanied by judgments of the event's practical significance and replicability."

I also see Daniel's (1998) views as being moderate, though they may tend a bit more toward the Carver (1978)/Rozeboom (1997) end of the continuum. Thus, the three articles do not include advocacy that the status quo is peachy-keen, and that no changes are warranted (a deficiency that will doubtless be corrected via additional commentaries).

My own views are fairly similar to those of McLean and Ernest (1998) and Daniel (1998). That is, on numerous occasions I certainly have pointed out the myriad problems with rampant misuse and misinterpretation of statistical tests.

However, I have never argued that statistical significance tests should be banned. If I felt these tests were intrinsically evil, as an editor of three journals, I necessarily would have written author guidelines proscribing these tests. And as an author I would also never report *p* values.

Instead, I generally find statistical tests to be largely irrelevant. Like Cohen (1994), I do not believe that *p* values evaluate the probability of what we want to know (i.e., the population). Rather, we assume the null hypothesis describes the population, and then evaluate the probability of the sample results (Thompson, 1996).

I am especially disinterested in statistical tests when what Cohen (1994) termed "nil" null hypotheses are used, particularly when testing reliability or validity coefficients. Daniel (1998) makes some excellent points here. We expect reliability and validity coefficients to be .7 or .8. As his table shows, with a *n* of 10 or 15, we will always attain statistical significance even for minimally acceptable reliability and validity coefficients, so what is the value of such tests with these or larger sample sizes? Abelson (1997) put the point fairly clearly:

> And when a reliability coefficient is declared to be nonzero, that is the ultimate in stupefyingly vacuous information. What we really want to know is whether an estimated reliability is .50'ish or .80'ish. (p. 121)

Thus, editorial policies of *Educational and Psychological Measurement* proscribe use of statistical testing of reliability and validity coefficients, if (and only if) "nil" nulls are used for this purpose.

I believe that evidence of result replicability is very important and is ignored by those many people who do not understand what statistical tests do (e.g., believe that their tests evaluate the probability of the population). Daniel (1998) at one point says, "Statistical significance simply indicates the probability that the null hypothesis is true in the population" (a view I do not accept), but says later that these tests answer the question, "*If* the sample represents the population, how likely is the obtained [sample] result?" (a view I do endorse).

Empirical studies consistently show that many researchers do not fully understand the logic of statistical tests (cf. Nelson, Rosenthal, & Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). Similarly, many textbooks teach misconceptions regarding these tests (Carver, 1978; Cohen, 1994).

More than anything else, I especially want to see authors always report effect sizes. I concur with the views of McLean and Ernest (1998), who noted that, "In reviewing the literature, the authors were unable to find an article that argued against the value of including some form of effect size or practical significance estimate in a research report." Kirk (1996) and Snyder and Lawson (1993) present helpful reviews of the many types of effect sizes that can be computed.

Regarding effect sizes, some (cf. Robinson & Levin, 1997) have argued that we must always first test statistical significance, and if results are statistically significant, "only if so: (2) effect size information should be provided" (Levin & Robinson, in press).

In Thompson (in press-b) I used a hypothetical to portray the consequences of this view. Two new proteins that suppress cancer metastasis and primary tumor growth in mice are discovered. Two hundred teams of researchers begin clinical trials with humans. Unfortunately, the 200 studies are underpowered, because the researchers slightly overestimate expected effects, or

perhaps because the researchers err too far in their fears of "over-powering" (Levin, 1997) their studies. Low and behold, all 200 studies yield noteworthy "moderate" effects for which $p_{CALCULATED}$ values are all .06.

> [A]m I to understand that these moderate effect sizes involving a pretty important criterion variable may not permissibly be discussed or even reported? . . . In the Thompson world, . . . [i]n this happy example, considerable direct replication evidence is available, so the noteworthy effect is interpreted even though none (zero, nada) of the 200 results is statistically significant. Thus, this is a world in which, in at least some cases, 'surely, God loves the .06 nearly as much as the .05' level of statistical significance (Rosnow & Rosenthal, 1989, p. 1277). (Thompson, in press-b)

Effect Size Reporting

Nix and Barnette (1998) cite others in suggesting that "studies today are more likely to report effect sizes," perhaps because the APA (1994) publication manual "encourages" (p. 18) such reports. However, McLean and Ernest (1998, emphasis in original) diametrically disagree, arguing that "encouraging" effect size reporting "has *not* appreciably affected actual reporting practices," and then cite five *empirical* studies corroborating their views.

Most regrettably, I believe that the pessimistic views of McLean and Ernest (1998) are correct. Indeed, let me cite five additional *empirical* studies of journal reporting practices that present similar findings (Keselman et al., in press; Lance & Vacha-Haase, 1998; Ness & Vacha-Haase, 1998; Nilsson & Vacha-Haase, 1998; Reetz & Vacha-Haase, 1998). In fact, Keselman et al. (in press) concluded that, "as anticipated, effect sizes were almost never reported along with *p*-values."

I have offered various reasons why the APA "encouragement" has been such a failure. First, an "encouragement" is too vague to enforce. Second, the APA policy

> presents a self-canceling mixed-message. To present an "encouragement" in the context of strict absolute standards regarding the esoterics of author note placement, pagination, and margins is to send the message, "these myriad requirements count, this encouragement doesn't." (Thompson, in press-b)

Of course, mindless adherence to old habits may also partly explain the glacial movement of the field, because "changing the beliefs and practices of a lifetime . . . naturally . . . provokes resistance" (Schmidt & Hunter, 1997, p. 49). As Rozeboom (1960) observed nearly 40 years ago, "the perceptual defenses of psychologists are particularly efficient when dealing with matters of methodology, and so the statistical folkways of a more primitive past continue to dominate the local scene" (p. 417).

It is my view (Thompson, 1998a; Vacha-Haase & Thompson, 1998) that most authors will simply not change their practices until editorial policies *require* them to do so. These three sets of authors cite three editorial policies (Heldref Foundation, 1997; Murphy, 1997; Thompson, 1994) requiring effect size reporting. Here are some additional editorial policies on this point. [Should *RESEARCH IN THE SCHOOLS* adopt such a policy? Hint. Hint.]

The editor of the *Journal of Consulting and Clinical Psychology* noted in passing that effect sizes are required in that journal, and furthermore that

> Evaluations of the outcomes of psychological treatments are favorably enhanced when the published report includes not only statistical significance and the *required* effect size but also a consideration of clinical significance. That is, . . . it is also important for the evaluator to consider the degree to which the outcomes are clinically significant (e.g., normative comparisons). . . . A treatment that produces a significant reduction in depressed mood must also be examined to determine whether the reduction moved participants from within to outside the defining boundary of scores for depression. (Kendall, 1997, p. 3, emphasis added)

The editor of the *Journal of Educational Psychology* called for "the provision of both strength-of-relationship measures and 'sufficient statistics' (the latter to permit independent confirmation of a study's statistical findings, statistical power calculations, and access to relevant information for meta-analyses, among others)" (Levin, 1995, p. 3).

The editor of the *Journal of Family Psychology* argued that, "In addition, reporting clinical significance . . . as opposed to mere statistical significance would also make treatment research more relevant to practitioners" (Levant, 1992, p. 6). Finally, the editor of the *Journal of*

Experimental Psychology: Learning, Memory, and Cognition argued that

> In reporting results, authors should still provide measures of variability and address the issue of the generalizability and reliability of their empirical findings across people and materials. There are a number of acceptable ways to do this, including reporting MSEs and confidence intervals and, in case of within-subject or within-items designs, the number of people or items that show the effect in the reported direction. (Neeley, 1995, p. 261)

Highlights of the Three Articles

The three articles each had highlights that particularly appealed to me. For example, Nix and Barnette (1998) present a nice albeit short review of the controversies between Fisher as against Neyman and Pearson, which were never effectively resolved (the consequence of this failed resolution being the hodge-podge of practices we see today). I very much liked their statement, "The *p* value tells us nothing about the magnitude of significance nor does it tell us anything about the probability of replication of a study." As I have noted elsewhere,

> 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 $p_{CALCULATED}$, and 100 studies with the same single effect size could each have 100 different values for $p_{CALCULATED}$. (Thompson, in press-a)

Daniel (1998) does a nice job of presenting older quotations to illustrate that we have been haunted by these controversies virtually since the inception of statistical tests. I particularly liked his citation of Berkson, arguing in 1938 that testing significance when the *n* is 200,000 is not very enlightening!

Daniel's (1998) review of editorial policies and how they are applied was also informative. He emphasizes a point that some authors do not appreciate: editors will not accept articles that violate their published editorial policies, so prudent authors must take these policies seriously. I find myself in general agreement with Daniel's (1998) very specific recommendations for improving our scholarship.

As regards McLean and Ernest (1998), I very much appreciated their recognition that science is subjective and that statistical tests cannot make it otherwise (Thompson, in press-c). I also very much liked their treatment of the "language controversy."

McLean and Ernest (1998) prefer to keep statistical tests within the researcher's arsenal but are more than willing to provide both effect size and replicability evidence of one or more sorts. I am somewhat less interested than they in the results of statistical tests, but science will move forward to the extent that the latter two issues are finally seriously considered within our inquiry.

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-141). Mahwah, NJ: Erlbaum.

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.

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.

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools, 5*(2) 23-32.

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

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

Harris, R. J. (1997). Reforming significance testing via three-valued logic. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 145-174). Mahwah, NJ: Erlbaum.

Heldref Foundation. (1997). Guidelines for contributors. *Journal of Experimental Education, 65*, 95-96.

Kendall, P. C. (1997). Editorial. *Journal of Consulting and Clinical Psychology*, *15*, 3-5.

Keselman, H. J., Huberty, C. J, Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (in press). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research.*

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

Lance, T., & Vacha-Haase, T. (1998, August). *The Counseling Psychologist: Trends and usages of statistical significance testing*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Levant, R. F. (1992). Editorial. *Journal of Family Psychology*, *6*, 3-9.

Levin, J. R. (1995). Editorial: Journal alert! *Journal of Educational Psychology*, *87*, 3-4.

Levin, J. R. (1997). Overcoming feelings of powerlessness in "aging" researchers: A primer on statistical power in analysis of variance designs. *Psychology and Aging*, *12*, 84-106.

Levin, J. R. (1998). To test or not to test H_0? *Educational and Psychological Measurement, 58*, 311-331.

Levin, J. R., & Robinson, D. H. (in press). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review*.

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools*, *5*(2)15-22.

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

Neeley, J. H. (1995). Editorial. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 261.

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

Ness, C., & Vacha-Haase, T. (1998, August). *Statistical significance reporting: Current trends and usages within Professional Psychology: Research and Practice*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Nilsson, J., & Vacha-Haase, T. (1998, August). *A review of statistical significance reporting in the Journal of Counseling Psychology*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: ban or abandon. A review of null hypothesis significance testing. *Research in the Schools*, *5*(2)3-14.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Reetz, D., & Vacha-Haase, T. (1998, August). *Trends and usages of statistical significance testing in adult development and aging research: A review of Psychology and Aging*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

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

Rosenthal, R. & Gaito, J. (1963). The interpretation of level 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.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335-392). Mahwah, NJ: Erlbaum.

Schmidt, F. (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.

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

Thompson, B. (Guest Ed.). (1993). Special issue on statistical significance testing, with comments from various journal editors. *Journal of Experimental Education, 61*(4).

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

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

Thompson, B. (1998a, April). *Five methodology errors in educational research: The pantheon of statistical significance and other faux pas*. Invited address presented at the annual meeting of the American Educational Research Association, San Diego.

Thompson, B. (1998b). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53*, 799-800.

Thompson, B. (1998c). Review of *What if there were no significance tests?* by L. Harlow, S. Mulaik & J. Steiger (Eds.). *Educational and Psychological Measurement, 58*, 332-344.

Thompson, B. (in press-a). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology*. [Invited address presented at the 1997 annual meeting of the American Psychological Association, Chicago.]

Thompson, B. (in press-b). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review*.

Thompson, B. (in press-c). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. *Theory & Psychology*.

Vacha-Haase, T., & Thompson, B. (1998, August). *APA editorial policies regarding statistical significance and effect size: Glacial fields move inexorably (but glacially)*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

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.