RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 63-65

The Statistical Significance Controversy Is Definitely Not Over: A Rejoinder to Responses by Thompson, Knapp, and Levin

Larry G. Daniel
University of North Texas

A rejoinder is offered on the three reviews of Daniel's article (this issue) by Thompson, Knapp, and Levin. It is concluded that the controversy over statistical significance testing will no doubt continue. Nevertheless, the gradual movement of the field toward requiring additional information in the reporting of statistical results is viewed as evidence of a positive response to long-term criticisms of statistical significance testing.

In this rejoinder, I would like to (a) respond to the critiques of Bruce Thompson, Tom Knapp, and Joel Levin of my earlier article in this issue and (b) provide additional commentary as to the future direction of statistical significance testing.

Response to Three Critics

I would like to express my appreciation to the three respondents for their insightful observations and for their comments casting further light on the issues raised by the authors of the three articles appearing in this issue of the journal. Each of the respondents is a premier scholar whose contributions to the debates on statistical significance testing have been most useful as the issue has come to the forefront of methdological discussions in recent years. In their critiques of the three articles included in this issue, the three respondents have offered very useful discussions of the topic along with helpful references for those readers who might wish to explore the controversy further. My specific comments in relation to the points made by each respondent follow in the order in which they appear in this issue of the journal.

Bruce Thompson (1998) provides a nice framework for understanding the ongoing dialogue regarding statisti-cal significance testing. Thompson's reminder of the context of the current literature in which much of the controversy has developed is useful in understanding the issue. This serves as a good follow up to the historical perspective that I provided. As Thompson noted, I have shared a long association with him and his work (he has been a mentor, research collaborator, and fellow editor); hence, I was not surprised that he was in agreement with many of the points I had raised and that a number of the opinions he expressed were consistent with my own. Further, I appreciate his citing the newly revamped editorial policies of several journals in addition to those that I had mentioned, lending evidence to the importance of editorial policies in shaping practice related to the reporting of results of statistical significance tests (SSTs). Further, Thompson (1998) reiterated nicely my discussion on the inappropriateness of using SSTs for the reporting of nil hypotheses about validity and reliability coef-ficients.

I am sure that Tom Knapp (1998) anticipated that the other authors and I would be eager to respond to his list of our various "errors of commission and omission." Obviously, determining what constitutes a sin is at least somewhat dependent upon the particular book of faith to which one prescribes. Although I prefer a slightly different statistician's book of faith than the one Knapp uses, I would have to say I am guilty as charged on at least a few points. First, I appreciate Knapp's (1998) comment on the distinction between the obtained and hypothesized effect sizes, an issue that often gets lost in the discussions of issues of this type. Second, I did indeed omit Levin's (1998a) excellent review of the *What If* book (Harlow, Mulaik, & Steiger, 1997) from my original discussion. This review is noteworthy not only because of Levin's excellent review of the content of the various chapters of the book, but also due to the concise list of recommended statistical significance practices that Levin offers. Third, I did not specifically mention the chapter in the *What If* book by Abelson (1997), which as Knapp (1998) indi-cated, is one of the more tightly written defenses of statistical significance testing.

Now that I have duly confessed, I would like to make a few citations from my own statistical book of faith on a couple of Knapp's other points. First, Knapp (1998) commented that resampling techniques such as jackknife and bootstrap analyses do not provide evidence of result replicability. (Levin [1998b] levels somewhat different but similarly focused criticisms at these procedures.) Even though the developers of jackknife and bootstrap techniques may not have specifically mentioned the usefulness of these procedures in providing evidence of

replicability, the procedures do indeed create varied resamplings for which results may be recomputed many times over. Clearly, the replications of results from these resamplings are somewhat biased and do not replace actual replications of the results with independent samples, but in newer areas of research, biased estimates of result replication are definitely better than no estimates of replication at all.

Knapp (1998) also questions the usefulness of "what if" analyses in which the results of SSTs are referenced to variations in sample size. Although I appreciate Knapp's concern that sample size should be carefully considered prior to the initiation of a study, it is often useful to determine at what sample size a statistically significant result would have become statistically nonsignificant and at what point a statistically nonsignificant result would have become statistically significant. These findings may advise researchers in selecting samples for *future* studies.

Knapp (1998) also splits hairs over the definition of the null hypothesis, apparently hinting at Cohen's distinc-tion between null hypotheses in their most "general sense" and "the nil hypothesis" that states that "the effect size (ES) is 0" (Cohen, 1994, p. 1000). Although this is an important distinction, Cohen (1994) reminded us that "as almost universally used, the null in H_0 is taken to mean nil, zero" (p. 1000); hence, my use of this conventional definition. Similarly, Knapp (as well as Levin, 1998b), commented on the technicalities of my example compar-ing SSTs with an *n* of 62 versus an *n* of 302. My intent was not to suggest that the relationship between *p* and F is linear, but rather to show with a fixed effect that results that were not statistically significant given a particular sample size would be much more likely to be statistically significant given a larger sample size.

Levin (1998b), in his predictably amusing style, provided some excellent comments on the several papers and the controversy. His comments on "statistical testiness" are especially interesting. As Thompson (1998) noted, not all scholars will have totally positive opinions about editorial policies, such as the ones I prescribed, that encourage specific practices in the reporting of the results of SSTs. Here, Levin voices at least one oft-heard complaint leveled at such editorial policies, namely, that regulation of specific verbiage transforms editors from being scholarly gatekeepers to statistical police. Although I am an ardent supporter of academic freedom, I do feel that regulation of vocabulary so as to avoid miscommunication is essential, and, as an editor, I have with some frequency felt it necessary to correct authors' verbiage so as to enhance their clarity of communication. Without a doubt, the term "significant" constitutes one of the more significant (pun intended) instances of miscommunication in social science literature, especially among readers who may not be familiar with the logic underlying SSTs. And, even though, as Levin (1998b) suggested, the specific written context may sometimes disambiguate the use of the term "significant," I would prefer to require routine use of "statistically" before "significant" so as to avoid overlooking instances in which the term should have been modified thusly but was not.

I feel that Levin somewhat overstated my position on statistical significance testing when he suggested I advocated that "the research world will be a far better place when the hypothesis-testing devil is ousted by the effect-size angel." Although I would clearly acknowledge the heavenliness of effect size reporting, I do not see hypothesis testing as the devil, but rather as an oft-tormented, though well-intended, soul who needs the demon of misinterpretation exorcized from him. In fact, in this regard, my position is not extremely dislike the one stated by Levin: report both effect size estimates and results of SSTs, then allow the readers of the research report to draw their own conclusions about result importance.

Comments on the Future of Statistical Significance Testing

Contrary to Levin's hopeful assertion that perhaps one day soon the bickering over statistical significance testing will be quelled, I do not see that happening very soon. Rather, I agree with Thompson (1998) that the status quo regarding the use of statistical significance testing is far from "peachy keen." Unfortunately, the literature is still rife with studies in which authors have misused and misinterpreted SSTs. As long as this remains the case, the voices of reformers as well as defenders of statistical significance testing will continue to be loudly heard. The battle will continue to rage for some time to come with perhaps an occasional quietus as other important methodological issues emerge followed by rekindling of the flames of debate as thoughtful research-ers continue to see errors in the reporting of SSTs.

Despite the slowness of progress in reforming practice relative to statistical significance testing, it is encouraging to see that an increasing number of social science journals are adopting editorial policies that call for better reporting of the results of SSTs (Thompson, 1998) following the suggestions found in the APA manual (APA, 1994). The adoption and enforcement of stricter editorial policies regarding the reporting of the

results of statistical significance testing by an increasing number of social science journals will perhaps eventually move the field toward improved practice. At the recent annual meeting of the Mid-South Educational Research Asso-ciation, Jim McLean, Co-Editor of this journal held a session in which he solicited input from the association members regarding the journal's potential adoption of an editorial policy on statistical significance testing. As a session participant, I was pleased to see that the group overwhelmingly favored such a policy. I look forward to seeing how Jim and Co-Editor Alan Kaufman handle the input gathered during that session.

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would have to 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: Author.

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

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

Knapp, T. R. (1998). Comments on the statistical significance articles. *RESEARCH IN THE SCHOOLS*, *5*(2), 39-41.

Levin, J. R. (1998a). To test or not to test H_0? *Educational and Psychological Measurement*, *58*, 313-333.

Levin, J. R. (1998b). What if there were no more bickering about statistical significance tests? *RESEARCH IN THE SCHOOLS*, *5*(2), 43-53.

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *RESEARCH IN THE SCHOOLS*, *5*(2), 33-38.