RESEARCH IN THE SCHOOLS
1998, Vol. 5, No. 2, 39-41

Comments on the Statistical Significance Testing Articles

Thomas R. Knapp
The Ohio State University

This review assumes a middle-of-the-road position regarding the controversy. The author expresses that significance tests have their place, but generally prefers confidence intervals. His remarks concentrate on ten errors of commission or omission that, in his opinion, weaken the arguments. These possible errors include using the jackknife and bootstrap procedures for replicability purposes, omitting key references, misrepresenting the null hypothesis, omitting the weaknesses of confidence intervals, ignoring the difference between a hypothesized effect size and an obtained effect size, erroneously assuming a linear relationship between p and F, claiming Cohen chose power level arbitrarily, referring to the "reliability of a study," inferring that inferential statistics are primarily for experiments, and recommending "what if" analyses.

Since I take a middle-of-the-road position regarding the significance testing controversy (I think that significance tests have their place, I generally prefer confidence intervals, and I don't like meta-analysis!), I would like to concentrate my remarks on ten errors of commission or omission that in my opinion weaken the arguments in these otherwise thoughtful papers. In this article, the three articles under review are referred to as Daniel (1998), McLean and Ernest (1998), and Nix and Barnette (1998).

1. Each of the authors discusses something they call "internal replicability analysis." The term is apparently due to Thompson (1994), and it represents a misinterpretation of the work on the jackknife and the bootstrap in the statistical literature. I challenge all of the authors to find in that literature (e.g., Diaconis & Efron, 1983; Efron & Gong, 1983; Mooney & Duval, 1993; Mosteller & Tukey, 1977) any reference to either approach providing evidence for the replicability of a finding. They are simply procedures for estimating sampling error without making the traditional parametric assumptions. The confusion may arise from the fact that both require the creation of several replications of the statistic of principal interest (the jackknife by "re-sampling" the sample data without replacement; the bootstrap by "re-sampling" the data with replacement).

2. None of the authors cite the article by Abelson (1997), and two of the authors (McLean and Ernest (1998) and Nix and Barnette (1998)) do not even cite the book on the significance testing controversy (Harlow, Mulaik, & Steiger, 1997) in which that article appears. It is the best defense of the use of significance tests I have ever read. Since the controversy has been going on for many years it is impossible to cite every relevant source, but McLean and Ernest (1998) don't even cite Schmidt (1996), the most vocal critic of significance tests and strongest advocate of meta-analysis. Daniel (1998) cites Thompson's (1998) review of the Harlow et al. compendium, but does not cite Levin's (1998) review that appeared in the same source.

3. Two of the authors make mistakes when discussing what a null hypothesis is. Daniel (1998) gives an example where the null hypothesis is said to be: *r* (the sample *r*) is equal to zero, and claims that "by definition" a test of significance tests the probability that a null hypothesis is true (the latter is OK in Bayesian analysis but not in classical inference). Both Daniel (1998) and Nix and Barnette (1998) refer to the null hypothesis as the hypothesis of no relationship or no difference; no, it is the hypothesis that is tested, and it need not have zero in it anyplace.

4. None of the authors point out the weaknesses of confidence intervals or how they can be misinterpreted just as seriously as significance tests. For example, it is not uncommon to see statements such as "the probability is .95 that the population correlation is between a and b." A population correlation doesn't have a probability and it is not "between" anything; it is a fixed, usually unknown, parameter that may be bracketed or covered by a particular confidence interval, but it doesn't vary.

Thomas R. Knapp is a professor of nursing and education at The Ohio State University. Correspondence regarding this article should be addressed to Thomas R. Knapp, College of Nursing, The Ohio State University, 1585 Neil Avenue, Columbus, OH 43210-1289 or send e-mail to knapp.5@osu.edu.

5. None of the authors make sufficiently explicit the necessary distinction between a hypothesized effect size and an obtained effect size. It is the former that is relevant in determining an appropriate sample size; it is the latter that provides an indication of the "practical significance" of a sample result and around which a confidence interval can be constructed. Cohen (1988) at least tried to differentiate the two when he put the subscript s on the d for the obtained effect size. Some of the confusion in the significance testing controversy could be avoided if we had different terms for those two kinds of "effect sizes." (A similar confusion has arisen recently regarding prospective and retrospective power – see Zumbo & Hubley, 1998.)

6. Daniel (1998) claims that a df of 300 for an ANOVA error term is five times more likely to produce a statistically significant difference than a df of 60. That's not true; the relationship between p and F is not linear.

7. McLean and Ernest (1998) claim that Cohen (1988) recommended a power of .80 as arbitrarily as Fisher recommended an alpha of .05. That's not fair. He (Cohen) argued there, and elsewhere, that Type I errors are generally more serious than Type II errors and therefore beta (= 1 - power) can be chosen to be considerably larger than alpha.

8. Nix and Barnette (1998) refer to "the reliability of the study." There is no such thing as the reliability of a study. Measuring instruments have varying degrees of reliability (I think the claim by Daniel (1998), and others, that reliability pertains to scores, not instruments, is much ado about nothing); statistics have varying degrees of reliability, in the sense of sampling error; studies do not.

9. Nix and Barnette (1998) also seem to suggest that inferential statistics in general and significance testing in particular are primarily relevant for experiments (given their several references to "treatments"). Statistical inference actually gets very complicated for experiments, since it is not clear what the population(s) of interest is (are). Experiments are almost never carried out on random samples, but all true experiments have random assignment. What inference is being made (from what to what) is a matter of no small confusion. (See the reaction by Levin, 1993 to Shaver, 1993 regarding this issue.)

10. Daniel (1998) advocates, as does Thompson, "what if" analyses (not to be confused with the "What if . . . ?" title of the Harlow book). Although such analyses are not wrong, they are unlikely to be very useful. Researchers have actual sample sizes and actual values for their statistics; speculating as to what might have happened if they had bigger or smaller sample sizes, or the population correlations had been bigger or smaller, or whatever, is the sort of thinking that should be gone through before a study is carried out, not after. (See Darlington, 1990, pp. 379-380 regarding this matter.)

But to end on a positive note, I commend Daniel (1998) for his point that a significance test tells you nothing about the representativeness of a sample; McLean and Ernest (1998) for their contention that significance tests (*and* confidence intervals?) aren't very important for huge sample sizes; and Nix and Barnette (1998) for bringing to the attention of the readers of this journal that there are both significance tests and confidence intervals available for multivariate analyses. Curiously, most of the controversy about significance testing has been confined to univariate and bivariate contexts.

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.

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

Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.

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.

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

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37*, 36-48.

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

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education*, *61*, 378-382.

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

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *RESEARCH IN THE SCHOOLS*, *5*(2), 15-22.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Newbury Park, CA: Sage.

Mosteller, F., &: Tukey, J.W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

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.

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

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

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

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

Zumbo, B. D., & Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician*, *47*, 385-388.