Toward Resolving the Significance Testing Debate: Electronic Publishing and Editorial Decision Making

David M. Lane and Miguel A. Quiñones

Rice University

This paper was presented at the 12th Annual Conference of the Society for Industrial and Organizational Psychology, April, 1997. St. Louis, MO.

Abstract

The debate over the use of statistical significance testing in the social sciences has heated up in recent years. This paper presents a brief review of the common criticisms of significance testing and argues that, as long as journal editors must choose among a number of manuscripts, significance testing provides useful information for making these choices. Electronic publishing is presented as a way of resolving the current impasse in this debate.

In a recent article, Schmidt (1996) portrays two visions of statistics and its role in psychological research. In the first, statistical significance testing is the primary means of data analysis. Although researchers are happily using the significance tests they learned in graduate school, these researchers have an inadequate understanding of the underlying logic of statistical inference. As a result, they use significance tests to answer questions that significance tests, in principle, cannot answer. This frequently leads to such serious misinterpretations of results that the accumulation of scientific knowledge is severely impede d. In the second vision, the shackles of significance testing have been thrown off and researchers rely primarily on interval estimation to analyze their results. Data from well-designed experiments are published without regard to significance testing and meta-analysis is used to combine the results across studies to reach valid conclusions.

Schmidt argues that significance testing has nothing valid to contribute to the analysis of data and that it should neither be taught in graduate school statistics courses nor included in published reports of scientific works. Although we are sympathetic to many of Schmidt's points, and agree with him that his second vision is one to be strived for, we believe Schmidt has overlooked an important characteristic of significance testing: significant results are more conclusive than nonsignificant results.

We begin this paper by summarizing the arguments of Schmidt and others (e.g., Cohen, 1994) against the use of significance tests. We then present the argument that significant results are more conclusive than nonsignificant results. This leads us to conclude that given the premium on journal space, editorial decisions will be and should be heavily influenced by the results of significance tests. This makes it very difficult to move toward Schmidt's second vision of the role of statistics. The proposed solution lies in a new model for publishing and editorial decision making based on electronic publishing on the World Wide Web.

Criticisms of Significance Testing

Numerous authors have pointed out errors that psychologists and other researchers have made interpreting and applying significance tests (see Chow, 1996, for a review). Among these errors are (1) accepting the null hypothesis for nonsignificant results, (2) interpreting the probability value as the probability the null hypothesis is false, (3) confusing statistical significance and effect size, and (4) interpreting the probability value as the probability of obtaining a significant outcome in a subsequent finding. Schmidt argued convincingly that the first of these errors, accepting the null hypothesis, has been the most detrimental to the advance of knowledge. As an example, Schmidt described how the failure of some studies to find significant correlations between employment tests and job performance led to numerous investigations of a phenomenon that is best explained as sampling error.

The crux of criticisms of significance testing is not that significance testing is misused, for that would only indicate that it should be used correctly. Instead, the argument is that significance testing has nothing of value to offer and that researchers use it because

they mistakenly believe that significance testing provides information that, in fact, it does not. For example, most researchers would like their statistical analysis to provide the probability that the null hypothesis is false. Instead, significance testing gives only the probability of the outcome (or a more extreme outcome) given that the null hypothesis is true. Critics of significance testing claim that most users of significance testing believe they have an objective way to determine whether or not a null hypothesis is true. However, in most if not all realistic experimental situations, the prior probability of the null hypotheses being true is essentially zero. To illustrate, suppose an experimenter were interested in comparing two methods for teaching subjects how to use a piece of computer software. It is not conceivable that the population difference between the two methods could be exactly zero. If the (population) mean time to perform a task after being trained with Method A is 3.6242541 minutes, is it conceivable that 3.6242541 minutes is also the mean time to perform after being trained with Method B? The argument, therefore, is that this whole to do about rejecting the null hypothesis is much ado about nothing since the null hypothesis is virtually always known to be false in the first place (cf. Cohen, 1994).

Significance Testing as Conclusiveness Testing

Despite these and other criticisms of significance testing, significance tests do make one important contribution: they indicate whether or not a set of experimental data is conclusive. Consider a “crucial” experiment in which competing theories make opposite predictions: Theory 1 predicts subjects in Condition A should outperform subjects in Condition B whereas Theory 2 predicts the opposite. Assume, for the sake of argument, that it is implausible that the two conditions result in exactly the same (population) level of performance. That leaves two possibilities: (1) Condition A > Condition B and (2) Condition A < Condition B.

Assume that an experimenter does a significance test and finds that the difference is statistically significant at the .01 level and that the mean for Condition A is greater than the mean for condition B. This finding allows the researcher to make a statement about the 99% confidence interval on the difference between population means. It is well known that if a statistic (the difference between sample means in this case) is significantly different from a hypothesized value (zero in this case) then the confidence interval associated with the significance test (99% confidence interval for .01 level) will not contain the hypothesized parameter. For this example, the significant outcome means that the 99% confidence interval on mean A -mean B does not contain zero. Instead, all values in the interval will be greater than zero. Thus all "plausible" values of mean A - mean B will be positive and the experimenter will be justified in concluding that mean A > mean B and therefore that Theory 1 is supported.

If the result had not been significant, then values on both sides of zero would have been included in the 99% confidence interval on mean A -mean B. This means that both theoretical outcomes are still plausible: mean A could be greater than mean B but mean B could be greater than mean A.

Although theories rarely live or die based on the result of a single study, a significant result certainly leads to a stronger conclusion than a nonsignificant result. Specifically, a significant result provides the basis for a researcher to draw a conclusion about the direction of an effect (see also Frick, 1996). Once this conclusion is reached, the experimenter may wish to try variations on the experiment to determine the boundary conditions of the effect or to see if the size of the effect depends on other factors. A nonsignificant result is an inconclusive result. As such, it does not support or confirm the null hypothesis. Instead, it fails to determine conclusively the direction of the effect.

As an aside, if significant results were called "conclusive results" then the propensity of researchers to accept the null hypothesis implicitly would be diminished. Instead of reporting "the difference between means was not significant" researchers would report "the direction of the difference between means was not determined conclusively."

Significance Testing and Editorial Decision Making

The negative consequences of using significance testing in editorial decision making have been discussed for some time (Bakan, 1966; Greenwald, 1975) . We believe the most serious problem is the incompatibility between significance testing and effect-size estimation. Specifically, if significant results are a criterion for publication, then published articles will contain inflated estimates of effect size (cf. Hedges, 1984; Lane & Dunlap, 1978). Since the power of psychological experiments is often relatively low (Cohen, 1962), this inflation can be substantial. For example, Lane and Dunlap found that when the true difference between two groups was 8 IQ points (SD=16) and alpha was set at .05, the observed mean difference between the two groups (each with n=10) was over 18 points when only significant results were considered. Naturally, as the alpha level was decreased (made more stringent), the amount of overestimation of the true difference rose dramatically. There have been some methods proposed for dealing with the bias inherent in our current system. Rosenthal (1979) presented a procedure for addressing the "file drawer" problem in meta-analytic research where only significant results are included. He showed that as the number of published studies increases, the probability of drawing incorrect conclusions by failing to include nonsignificant unpublished studies becomes trivial. However, Rosenthal's analysis does not speak to the issue of bias in estimating effect size. Hedges (1984) developed an analytical procedure for estimating effect size based on a set of estimates from a distribution truncated by including only significant results. Although Hedges's procedure is a major contribution, it is not a perfect solution. For example, complications arise when some but not all of the published articles report significant results. Hedges recommends that in such situations, the nonsignificant results be discarded and his procedure applied to the significant outcomes. Although this is a generally good solution, there are occasions in which this would result in an unacceptably large amount of information being lost. Moreover, complex situations such as one in which the probability of a paper being accepted is a continuous monotonic function of the probability level are difficult to accommodate. In short, no method for correcting for a bias can be quite as good as not having the bias in the first place.

The bias would be eliminated by basing editorial decisions solely on the basis of a paper's introduction and method section. However, there are two problems with this approach. First, as pointed out by Lane and Dunlap (1978), an experiment based on an unconventional theoretical perspective would not be very interesting if the data contradicted the theory. Second, studies with conclusive results will (and should) be preferred to studies with inconclusive results. Consider an editor who, due to limited (and expensive) journal space, can only accept one of two papers being considered. Both papers address equally-important topics using equally rigorous methods. Paper 1 seeks to determine the relative effectiveness Conditions A and B while Paper 2 seeks to determine the relative effectiveness of Conditions C and D. In Paper 1, Condition A is significantly better than Condition B, allowing the conclusion that, in the population, Condition A is better than Condition B. In paper 2, Condition C is better, but not significantly better, than Condition D. Paper 2, therefore, is unable to conclude which condition is better in the population. The editor is faced with choosing between a conclusive experiment and an inconclusive one. There seems little doubt that the conclusive experiment should have a higher priority.

Although only one would be accepted, both of the above papers are worthy of publication in the sense that they contain contributions to the field. A scientist doing a meta-analysis of the difference between Conditions C and D would certainly be interested in the results of Paper 2 even though the results of that paper do not stand on their own. Moreover, if other experiments are conducted comparing Conditions C and D and only the significant ones are published, the true difference between these conditions will be vastly overestimated (Hedges, 1984; Lane & Dunlap, 1978). Nonetheless, given the premium on journal space, the first paper would have priority over the second. Editorial decision making must involve judging the contribution of one paper relative to the contributions of other papers vying for publication. Since a major aspect of a paper's contribution is the conclusiveness of its results, statistical significance necessarily plays a critical role in the decision process.

Electronic Publishing and Significance Testing

A recent article in Science reports on the huge explosion in electronic publishing in the physical sciences (Taubes, 1996): At the end of 1995, over 100 peer-reviewed science journals were available over the internet. Some of these journals use electronic publishing as a supplement to their regular paper publication. However, an increasing number of journals are becoming strictly on- line publications. For example, the psychology journal Psycoloquy has been around for several years and is only available in electronic form. A critical difference between electronic and paper publications is that the marginal cost of an electronically-published article is negligible. It is known from microeconomic theory that a firm should continue to increase production as long as the marginal revenue is greater than the marginal cost. In the present context, this means that a paper should be published as long as an article makes a positive contribution to the field. Therefore, unlike the present publication system where the contributions of papers are judged relative to contributions of other papers, electronic publishing allows papers to be judged on their own merit. A well- designed study producing inconclusive results makes a positive contribution to the field and should be published. Since the cost

of making this information available to the research community is negligible, it is hard to justify keeping the information from being disseminated.

The policy of ignoring the outcome of significance tests would be of great benefit. Researchers could use significance tests as a short hand for whether a confidence interval contains zero. However, they would be encouraged to refer to these as "conclusiveness" tests thus avoiding two potential misuses of significance testing: (a) using significance testing as a measure of effect size and (b) accepting the null hypothesis when it is not rejected.

Naturally, researchers interested in estimating effect size would find the elimination of significance testing as a criterion for publication highly desirable. Schmidt's concern that significance testing is hindering the accumulation of scientific knowledge would be addressed.

There are a number of other benefits of on-line publications. One important benefit is speed of publication. By eliminating the production phase, it is possible to go from submission to publication in a manner of weeks rather than months. Another advantage is the ability to search the journal using key words and phrases. It is also possible for articles to provide links to related articles or data throughout the text or in the reference section. Authors could provide a link to the raw data used in the study. This would allow scientists to double check the work and replicate the analyses.

Several objections could be raised to the proposition that journals should be published electronically and that these electronic journals should not use significance testing as a criterion for publication. One objection is based on what Greenwald (1975) calls the cultural truism that "... incompetence is more likely to lead to erroneous nonsignificant, 'negative,' or null results." (p. 2). In refuting this "cultural truism," Greenwald acknowledges that incompetence can have the effect of introducing noise into the data. For example, an incompetent experimenter could increase error variance by making random errors in data transcribing, by running the experiment in an environment with distracting noise, or by inaccurately placing electrodes. However, as Greenwald points out, other more common types of incompetence result in systematic errors and thus a tendency to falsely reject the null hypothesis. Examples include demand characteristics, nonrandom sampling, invalid or contaminated manipulations, and apparatus malfunctions.

A second objection to eliminating significance testing as a criterion for publication is that so many studies would be published that there would be an information overload. We believe a policy that rejects valid studies simply because their publication would make it more difficult for researchers to stay current would be a mistake. Although publishing more papers would certainly require some adjustments such as increasing the number of articles that are literature reviews and/or meta-analyses, we believe the problems would be relatively minor.

The single largest barrier to electronic publishing is probably the attitude of the academic community itself. It is clear that articles published in new on-line journals would not carry the same weight as those published in more established print journals. It is probably a matter of time, however, before attitudes towards on-line publications change and the major journals are published electronically. Electronic publishing will need to maintain a high level of quality control and editorial oversight. The major difference is that the quality control will be focused more on the theoretical justification and experimental methods and less on the outcome of significance tests.

Conclusion

All indications are that the debate over significance testing will continue for some time. Our position is that until all meritorious studies can be published, the present system for deciding which studies are more worthy of publication is necessary. Significance testing allows one to make statements about the conclusiveness of results and, therefore, in spite of the adverse consequences of doing so, significance testing should continue to be used as an important criterion in editorial decision making.

We propose that electronic publishing may provide the answer to the current dilemma. By decreasing the marginal costs of publication, practically all theoretically sound and well-designed studies can be published without regard to the statistical significance of the results. This should decrease the emphasis on null-hypothesis testing as well as increase the validity of meta-analyses. Given the current rate of growth on the internet, this vision may soon be practical.

References

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 432-437.

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

Chow, S. L. (1996). *Statistical significance: Rationale, validity, and utility*. Thousand Oaks, CA: Sage.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145-153 .

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

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

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

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*, 61-85.

Lane, D. M. & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*, 107-112.

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

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

Taubes, G. (1996). Science journals go wired. *Science*, 271, 764-766.