DOCUMENT RESUME

ED 344 905 TM 018 225

AUTHOR Shaver, James P.
TITLE What Statistical Significance Testing Is, and What It Is Not.
PUB DATE Apr 92
NOTE 43p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Research; Evaluation Problems; Hypothesis Testing; Probability; Psychological Studies; *Research Design; Research Problems; *Sample Size; *Statistical Significance; Test Validity
IDENTIFIERS *Null Hypothesis; *Randomization (Statistics); Research Replication

ABSTRACT

A test of statistical significance is a procedure for determining how likely a result is assuming a null hypothesis to be true with randomization and a sample of size n (the given size in the study). Randomization, which refers to random sampling and random assignment, is important because it ensures the independence of observations, but it does not guarantee independence beyond the initial sample selection. A test of statistical significance provides a statement of probability of occurrence in the long run, with repeated random sampling under the null hypothesis, but provides no basis for a conclusion about the probability that a particular result is attributable to chance. A test of statistical significance also does not indicate the probability that the null hypothesis is true or false and does not indicate whether a treatment being studied had an effect. Statistical significance indicates neither the magnitude nor the importance of a result, and is no indication of the probability that a result would be obtained on study replication. Although tests of statistical significance yield little valid information for questions of interest in most educational research, use and misuse of such tests remain common for a variety of reasons. Researchers should be encouraged to minimize statistical significance tests and to state expectations for quantitative results as critical effect sizes. There is a 58-item list of references. (SLD)

WHAT STATISTICAL SIGNIFICANCE TESTING IS,

AND WHAT IT IS NOT

James P. Shaver

Utah State University

Paper prepared for a symposium, Significance Testing in a Post-positivistic Era: Some Proposed Alternatives, with Comments from Journal Editors, at the annual meeting of the American Educational Research Association, San Francisco, April 22, 1992.

Kenneth E. Bell, David G. Gibson, Perry J. Sailor, and Matt Taylor provided helpful comments on the paper.

The use of tests of statistical significance in educational and psychological research has been under attack for over 30 years. Among the critics have been Skinner (1956), Bakan (1967), Meehl (1967), various authors in Morrison and Henkel (1970), Seeman (1973), Signorelli (1974), and Cronbach (1975). In 1978, Carver's excellent critique, "The Case Against Statistical Significance Testing," was published. I touched on the matter as it related to the lack of productivity of educational research in 1979, participated in an American Educational Research Association symposium in 1980 with a title similar to the one for which this paper was prepared--"Tests of Statistical Significance: Readdressing Their Role", and then wrote a two-part article cautioning educational practitioners about misinterpretations of statistical significance (Shaver, 1985a, b).*

In 1978, Carver (p. 379) noted that all of the criticisms of tests of statistical significance appeared to have had little effect. The situation has not changed since then. A quick perusal of educational research journals, educational and psychological statistics textbooks, and doctoral dissertations will confirm that tests of statistical significance continue to dominate the interpretation of quantitative data in educational research. Surely one characteristic of statistical significance testing is that it is an enduring--in the face of the devastating criticism, perhaps it would be better to say, relentless--phenomenon in educational and psychological research.

The thrust of this paper, like so many written before it, is that the dominance of statistical significance testing is dysfunctional, because such

* Much of this paper is a story told before. That has presented a quandary in regard to how extensively to develop various concepts and to cite supporting sources for ideas that seem well established, if not well accepted. I have probably been both over-frugal and excessive on both accounts at different points in the paper.

tests do not provide the information that many researchers assume they do. Statistical significance testing also diverts attention and energy from more appropriate strategies, such as replication and attention to the practical or theoretical significance of results. In the hope that the accumulation of criticism will have an effect, I respond again in this paper to the question of what statistical significance testing is and what it is not. Possible reasons for the persistence of statistical significance testing are also discussed briefly, and proposals are presented for action by journal editors to moderate the negative effects of statistical significance testing, if not eradicate their inappropriate use.

What Statistical Significance Testing Is

A test of statistical significance is, at its very simplest in the dominant Fisherian model of hypothesis testing, a procedure for determining how likely a result is assuming a null hypothesis to be true. Somewhat more precisely, our commonly used tests of statistical significance (*z*-ratios, *t*-ratios, and *F*-ratios, such as in the analysis of variance or covariance) are procedures for determining the probability (usually at a prespecified level called alpha) of a result under the null hypothesis (assuming the null hypothesis to be true) with randomization* and a sample of size *n* (i.e., the sample size used in the study).

Individual elements of that statement are important, although often overlooked. First, the result of a test of statistical significance is a probability statement, often expressed as a dichotomy in terms of whether the

* I use randomization to include both random sampling and random assignment, although some authors use randomization to refer only to assignment. Much of the discussion that follows is in terms of random sampling; but, as I point out later, random assignment also meets the randomness assumption.

probability was less or greater than the alpha level. Second, the test is based on the assumption that the null hypothesis is true. That is, the theoretical sampling distributions against which results are compared (the normal distribution, the *t*-distributions, the *F*-distributions, the chi-square distributions) are generated by assuming that sampling occurs from a population, or populations, in which the null hypothesis is true. Third, despite some claims to the contrary (e.g., Thompson, 1987), randomization is a fundamental assumption underlying the use of these tests of statistical significance. Fourth, sample size is a crucial consideration, because the statistical significance of a result will depend on the number of cases on which it is based. Each of these elements will be alluded to in the discussion that follows.

Randomness as an Assumption

As Glass and Hopkins (1984) stated, "Inferential statistics is based on the assumption of random sampling from populations" (p. 177). Elsewhere, they refer to "random samples" as one of the "building blocks for hypothesis testing" (p. 202), and they specify random sampling as an essential assumption for the use of the one-sample *z*-test for a mean (p. 205), the *t*-tests for two independent means (p. 231) and for the difference between means from correlated observations (p. 241), and the *F*-ratios for differences between variances (p. 261) and in the analysis of variance (e.g., pp. 342, 445).

Randomization is important because it helps to ensure the independence of observations (or, equivalently, errors; Glass & Hopkins, 1984, p. 350). Despite what is commonly assumed, however, randomness does not guarantee independence beyond the initial sample selection. For example, observations almost certainly will not be independent when treatments have been delivered

to subjects in a group setting, as is common in educational research.

In addition, randomization is essential to the typical tests of statistical significance. Randomness (i.e., random error) is the basis for the sampling distributions against which results are compared. Use of, for example, a t-distribution to answer the question, "How likely is this particular result under the null hypothesis?" will not yield a meaningful probability statement if the sample is not random. Repeated random sampling (or assignment) yields known sampling distributions. Nonrandom sampling does not, nor does the comparison of a nonrandom sample to a randomly generated sampling distribution provide a valid statement of probability of occurrence.

The indispensability of randomness may be more evident when the question being addressed in tests of statistical significance is stated as, how representative is this sample (sample statistic) of the population (population parameter) as specified in the null hypothesis? Without randomness, that question cannot be answered validly using the common tests of statistical significance. As Glass and Hopkins (1984) put it:

> The method of random selection of samples will ensure, within a certain known margin of error, representativeness of the samples and hence will permit establishing limits within which the parameters are expected to lie with a particular probability.
>
> The ability to estimate the degree of error due to chance (sampling error) is an important feature of a random sample [emphasis in the original] . . . . It is not possible to estimate the error with accidental sampling and many other sampling strategies since they contain unknown types and degrees of bias in addition to sampling error. (p. 177)

In that context, I found it baffling that Thompson (1987) would assert that "significance testing imposes a restriction that samples must be representative of a population, but does not mandate that this end must be realized through random sampling" (pp. 8-9), and then go on to discuss "comparing known sample characteristics with known population characteristics to build some warrant for an assumption of representativeness" (p. 9). The description of sample characteristics in order to allow generalization to populations from which a random sample was not drawn is an important, and often neglected, element of research reporting (Shaver & Norton, 1980a, b). In fact, such description of sample characteristics is crucial, even if a random sample was used, to assist readers in making generalizations—both because the sample may not have been drawn from a population in which a research user is directly interested and, equally important, because a random sample may not represent well the population from which it was drawn.

Such descriptions are not, however, a substitute for random sampling; the purpose of randomness is not to ensure representativeness (if that could be done, there would be no need for an inferential test), but to allow the specification of the probability that a sample came from a population with an hypothesized parameter (or, conversely, to estimate a range of probable values for a parameter). In short, random sampling addresses solely the representativeness of samples in the long run; it does not ensure that all of the characteristics of a particular sample, including the dependent variable(s) under investigation, will be the same as those of the population, only that (whether assessed or not) they will differ only by chance from the population characteristics. Of course, this also means that in conventional significance testing with random sampling from a population in which the null

hypothesis is true and with alpha set at .05, 5% of the time the researcher will incorrectly conclude that the sample did not come from the population specified in the null hypothesis; however, a conclusion that the sample was not representative of the specified population (with α as the criterion) would be correct.

In essence, the mistake is in viewing randomization as an outcome (i.e., representativeness) rather than as a process (i.e., sampling in which every member of the population has an equal chance of being selected for the sample). This error is not uncommon, and can even be found in statistics books. For example, Ferguson and Takane (1989) provided an example of the use of chi-square "to test the representativeness of a sample where certain population values are known" (p. 218). They analyzed a set of data composed of 200 individuals drawn (the process is not specified) from the population of Montreal. The difference between population and sample frequencies for three levels of national origin--French, English, and other--was statistically significant at the .01 level. Ferguson and Takane concluded, erroneously, "that the sample . . . cannot be considered a random sample" (p. 219). Of course, the question of randomness is not a matter of a chi-square goodness of fit test, but of the process by which the sample was drawn. Had they said that the sample could not be considered a "representative" sample, their logic would have been correct, although statistical significance has dubious validity as the criterion for such a decision.

Random assignment. The term randomization is used somewhat ambiguously in discussions of experimental design and tests of statistical significance. Some (e.g., Hays, 1973, p. 562) use randomization to refer generally to the application of random processes in designing experimental studies,

encompassing both random sampling and random assignment. On the other hand, there are those who use randomization to refer only to random assignment to treatments, (e.g., Ferguson & Takane, 1989, p. 245; Winer, Brown, & Michels, 1991, pp. 7-8). Focusing on random sampling, and not discussing random assignment, is also common. That is the case with Glass and Hopkins (1984), who discuss tests of statistical significance in terms of random sampling but not random assignment to treatments.

Although random assignment is not common in educational research, it is more so than random sampling (Shaver & Norton, 1980a, b). As Berk and Brewer (1978) pointed out, with random assignment, researchers can appropriately compare their results against the sampling distributions commonly used in tests of statistical significance (also see Hays, 1973, p. 562; Winer et al., 1991, p. 8). Whereas random sampling ensures chance sample differences from the source population on all characteristics, random assignment ensures that differences between the groups on all variables, assessed or not, are nonsystematic. Again, there is no assurance that the groups are not different on any important variable. In fact, a test of statistical significance may indicate, even after random assignment, that the groups are sufficiently different on the variable(s) under analysis that, following the logic of the inferential test, one should conclude that they did not come from the same population.

Repeated random assignment to groups from the same population will result in a sampling distribution of mean differences with a mean equal to zero. The z, t, or F distributions can be used in tests of statistical significance to determine the probability of a particular mean difference (or difference in some other statistic) under the null hypothesis of no

difference. Of course, with only random assignment, a test of statistical significance provides no basis for generalization to a specific population, although it can be regarded as addressing the question whether the groups under analysis can be regarded as samples from the same undefined population.

Just as random sampling will not ensure that a particular sample is representative of the population from which it is drawn, random assignment does not provide assurance that the resulting groups are identical with one another or, put alternatively, that the groups are equivalent splits from the same hypothetical infinite population (McHugh, 1964). Random sampling into treatment groups addresses both the estimation of population parameters and the likelihood of associations between treatment group membership and preexisting characteristics; random assignment addresses only the latter.

It should be noted that, without random assignment, an exact probability test of statistical significance can be based on a sampling distribution generated by randomly splitting the available sample into all possible combinations of the size of the groups in the study and computing the relevant statistic for each combination. The researcher can then ask, using that distribution, how likely it is that an obtained result would have occurred by chance (e.g., Berk & Brewer, 1978; Winch & Campbell, 1969). Such probability tests are rarely reported in the literature, however. Traditional tests of statistical significance are typically applied, often ignoring the assumption of randomization.

Violations of randomness. Unfortunately, statistics textbook authors tend to ignore the effects of violating the randomness assumption. For example, Glass and Hopkins (1984), who are explicit about the importance of random sampling if not random assignment, discuss the effects of violating the

assumptions of normal population distributions and homogeneous population variances, but not the effects of lack of randomness on the appropriateness of drawing a conclusion about a particular result using the theoretical sampling distribution. The effect of lack of randomness on the independence of scores, is, however, often mentioned (e.g., Glass & Hopkins, 1984, p. 353).

One reason that randomness is often ignored may be that the examination of the effects of violating that assumption is a formidable task because it involves all sample-population characteristics, not only the dependent variable as with the normality and homogeneity of variance assumptions. As I have pointed out earlier (Shaver, 1980):

> To enumerate every potentially relevant variable and specify its relationship to the dependent variable(s), being certain that no crucial variable was overlooked, in order to investigate the effects of nonrandomness on probability statements presents insuperable difficulties. (p. 6)

Nevertheless, the general conclusion that levels of randomness can be overlooked, as is common in the reporting of educational research, must be challenged. As Winer et al. (1991) stated in discussing analysis of variance assumptions:

> Violating the assumption of random sampling of elements from a population and random assignment of the elements to the treatments may totally invalidate any study, since <u>randomness provides the assurance that errors are independently distributed</u> [emphasis added] within and between treatment conditions and is also the mechanism by which bias is removed from treatment conditions. (p. 101)

However, their table summarizing the "consequences of violation of assumptions of the fixed-effects ANOVA" (p. 102) includes randomization only in terms of the independence of observations (errors), perhaps because the other consequences are unknown, and unknowable in practice.

An analogy. To sum up, the commonly used tests of statistical significance provide the researcher with limited information: How likely is this result, assuming the null hypothesis to be true and with randomization (random sampling and/or assignment) and a sample of size n? Without randomness, the result of the test of statistical significance is meaningless or, at best, its relevance to a statement of probability is indeterminate. Frequently in educational research, the researcher goes into a school or schools, obtains available groups (with neither random sampling or assignment), collects data—sometimes with, and sometimes without, a treatment—and then conducts tests of statistical significance. The results of such inferential tests are essentially meaningless, unless one is interested in comparisons to an abstract standard of probability as indicated by the question: What would be the probability of the obtained result if random samples had actually been drawn?

Consider an analogous situation: A person walks into a room and sees 10 coins lying on a table. He observes 8 heads and 2 tails, and wonders if the coins are biased. So he asks if this particular arrangement of coins is a likely chance occurrence. Having a statistics book handy, he turns to Pascal's triangle and finds that the probability of obtaining 8 out of 10 heads is 45/1024 or .044. Because that probability is less than the traditional .05 alpha level, he concludes that his result is not a likely chance occurrence under the null hypothesis of a 50-50 split in tails and

heads, and that he has evidence that the coins are biased. However, what he clearly should conclude is that if he has flipped each coin--or, alternatively, had flipped 1 coin 10 times--the probability of the particular result occurring by chance is less than 5%. But he has no evidence as to the bias in the particular set of coins because they were not, as far as he knows, flipped. That is, he does not know the process by which they arrived in their positions. The theoretical (binomial) distribution provides only an abstract standard of little relevance because the data were not produced in such a way as to meet a basic assumption for use of the distribution. (Describing the physical properties of the coins vis-a-vis biasedness would be a substitute for flipping and use of the binomial distribution, not proof that the binomial distribution was applicable.)

Just as application of the binomial distribution could not provide valid information about possible bias in the observed coins, so educational researchers who use nonrandomized groups cannot obtain valid information about the probability of a group difference under the null hypothesis using a common test of statistical significance. The t-distribution, for example, has no more relevance to differences between available groups than the binomial distribution does to the possible bias in coins found lying on a table.

What Statistical Significance Testing Is Not

A test of statistical significance used without randomization, then, does not yield valid information about the probability of a result under the null hypothesis. The following brief listing of what tests of statistical significance cannot do for the researcher is, therefore, based on the assumption that data come from a design that includes randomization--either random sampling or random assignment. As noted above, with randomization, a

test of statistical significance provides a researcher with information on the probability of a result assuming the null hypothesis to be true and given the sample size. On the other hand, a test of statistical significance does not provide information on a number of matters of interest to researchers, even though it is often presumed to do so.

What About This Sample?

A test of statistical significance provides a statement of probability of occurrence in the long run, with repeated random sampling (or assignment) under the null hypothesis. As Carver (1978) argued, it is a fantasy to believe that such a test speaks to whether a particular result is a chance occurrence. That is, a test of significance provides the probability of a result occurring by chance in the long run under the null hypothesis with random sampling and sample size n; it provides no basis for a conclusion about the probability that a particular result is attributable to chance.

Even with random samples drawn from a population in which the null hypothesis is true and in which scores are distributed according to the assumptions for the statistical model, with alpha set at .05, 5% of the time the researcher will conclude that a result is not a likely occurrence under the null hypothesis, thus making a Type I error. But the researcher has no way of knowing for any particular result whether that error is being made or if the sample was drawn from a population in which the null hypothesis was not true. That is why, according to Tukey (1969), R. A. Fisher's "standard of firm knowledge was not one very extremely significant result, but rather the ability to repeatedly get results significant at 5%" (p. 85). Replication is essential to confidence in the reliability (reproducibility) of a result, as well as to conclusions about generalizability (external validity) (e.g.,

Campbell & Jackson, 1979).

What About H_0?

A test of statistical significance does not indicate the probability that the null hypothesis is true or false. It provides the researcher with information in regard to the likelihood of a result given that the null hypothesis is true; it does not indicate the likelihood that the null hypothesis is true given a particular result. Carver (1978) and J. Cohen (1990) are among those who have cautioned against that fallacy.

Unfortunately, authors of statistics books and research reports frequently make statements about rejecting the null hypothesis based on one statistically significant result. That is too absolute a conclusion. If any inference about the null hypothesis is to be drawn from one test of statistical significance, it should be stated in terms of evidence for the plausibility of the null hypothesis, not an absolute rejection. Even though the "rejection" decision may be set implicitly in a probabilistic context--with, e.g., a .05 chance of a Type I error--that qualification is, typically, quickly ignored.

Conversely, a test of statistical significance also does not provide information on the probability that an alternative hypothesis is true or false (see, e.g., Carver, 1978). The sampling distribution is based on the null hypothesis; so no evidence is provided as to the likelihood of the result occurring under alternative hypotheses.

What About Treatment Effects?

One of the more egregious errors is to conclude that a test of statistical significance indicates whether a treatment being studied had an effect. Clearly, the test of statistical significance addresses only the

simple question whether a result is a likely occurrence under the null hypothesis with randomization and a sample of size n. At most (as noted above), a statistically significant result has to do with the probability of the result in the long run (i.e., over repeated samples), not with whether the particular result did or did not occur under the null hypothesis.

It seems terribly obvious that a test of statistical significance does not speak directly to causality. Even if a researcher were willing to conclude, following a test of statistical significance, that a low probability result did not come from a population in which the null hypothesis is true (i.e., the result was not a chance occurrence under H_0), and did not make a Type I error in doing so, the test of statistical significance provides no evidence as to the cause of the result. That is a matter of design, not of statistical inference. Not even the selection threat to internal validity is perfectly controlled by random sampling or assignment; simply by chance pretreatment groups can be significantly (statistically and/or practically) different on one or more relevant variables.

The fallacy of concluding that a statistically significant result indicates a treatment effect (as often seen, for example, in statements such as, "the statistically significant difference between means indicates that Treatment A was more effective than Treatment B") is likely perpetuated by statistics books in which it is suggested that tests of statistical significance will address questions such as: "Is the treatment effective? Does Drug X reduce hyperactivity more than a placebo? Does anxiety level influence test performance?" (Glass & Hopkins, 1984, p. 230).

Statements such as, "the mean practice effect was highly significant" (Glass & Hopkins, 1984, p. 242), following a t-test comparing pre- and

posttest means, can mislead as well, as can a statement that a confidence interval for a difference between means indicates the range of values that encompass "the true treatment effect ($\mu_1 - \mu_2$)" (p. 236). The ambiguity of the term effect as used in inferential statistics, especially analysis of variance, to refer to a comparison (e.g., a "main effect") does not help. As J. Cohen and P. Cohen (1983) pointed out, the "causal implication of the term effect" can obscure the fact that "causal interpretations are never warranted by statistical results, but require logical and substantive bases" (p. 210).

Nontextbook discussions to inform researchers about inferential statistics can misinform, as well. For example, J. Cohen (1990) commented that "everyone knows that . . . all [statistical significance] means is that the effect is not nil . . ." (p. 1307). And, in a generally sound piece, Berk and Brewer (1978) said: "If this null hypothesis [$\mu_1 = \mu_2$] is rejected, it can be concluded that presence or absence of the specific diploma treatment contributes to group differences in income" (p. 209). Such statements create and perpetuate an erroneous view of the relationship between tests of statistical significance and causality.

What About Magnitude and Importance?

Despite frequent conclusions to the contrary in research reports (as noted, e.g., by Bracey, 1991, and Harcum, 1989), statistical significance indicates neither the magnitude nor the importance of a result. Statistical significance is only information in regard to the probability of a result under the null hypothesis with randomization and a sample size of n.

Sample size is, of course, a primary concern in this particular instance of what a statistical significance test is not. For example, with $n = 10$ and $\alpha = .05$ for a nondirectional test, a correlation of .63 is statistically

significant; with $n = 50$, an $r = .28$ is needed; with $n = 100$, an $r = .20$; with $n = 500$, an $r = .09$; with $n = 1,000$, an $r = .06$; and, with $n = 10,000$, an $r = .02$ is statistically significant at the .05 level (Glass & Hopkins, 1984, p. 549). Alternatively, with a standard deviation of 10 and $n = 20$, a difference of 9.4 between 2 independent means is necessary for statistical significance at the .05 level in a nondirectional test; with $n = 100$, a difference of only 4.0 is required, and with $n = 1000$, a difference of only 1.2 is required.

Obviously, very small and trivial results as well as important ones may be statistically significant. As Meehl (1967), along with a host of other writers, has pointed out, with a large enough sample and reliable assessment, practically every association will be statistically significant. Conversely, with a very small sample, very few results will be statistically significant. Therefore, to know only whether a result is statistically significant tells one virtually nothing about the magnitude or importance of the result.

It is so commonly stressed that the statistical significance of results is directly a function of sample size that one can only wonder at the number of articles in which results are either interpreted as important because of statistical significance or in which the probability level appears to be taken as an indication of magnitude, as suggested by the use of terms such as "highly significant" when the probability is .01 or less. Even statistics textbook authors make the latter mistake, as in Glass and Hopkins' (1984) references to a "mean practice effect [that] was highly significant" (p. 242) and a "multiple correlation coefficient [that] is highly significant" (p. 314) when the probabilities were .001 or less.

Effect sizes. Statistical probability is not, then, a useful indicator of magnitude of a result because it is dependent on sample size. This

deficiency became especially clear in efforts to prepare quantitative research-literature summaries. Glass (1976) proposed that effect sizes*--metrics for the magnitude of results that are independent of sample size and scale of measurement--be used in reporting results.

Effect sizes, too, may be misinterpreted, however. Despite J. Cohen's (1988, p. 12-13) cautions to the contrary, researchers have ignored the arbitrariness of his conventions for low, medium, and large effect sizes (e.g., .2, .5, and .8, respectively, for standardized mean differences). Yet, as has been made clear by a number of authors (e.g., Glass, McGaw, & Smith, 1981, p. 104; Shaver, 1985b, 1991), an effect size of 1 or larger may reflect a trivial result. The dependent variable may lack value (benefit), the construct or characteristic may not have been validly assessed (Messick, 1989), or the result may be too costly to produce or its reliability may be in doubt. Substituting sanctified effect size conventions for the sanctified .05 level of statistical significance is not progress.

Power analysis. One reaction to the relationship between sample size and statistical significance has been the call for statistical power analysis. J. Cohen (1988), among others, has indicted the low power--that is, the probability of obtaining a statistically significant result if there is a difference in the population--of much psychological and educational research due to the small sample sizes typically used and the small to moderate effect sizes that are common.

In conducting a power analysis, the population value is not known (otherwise, a test of statistical significance would be irrelevant); an effect

* Result size would be a better term to avoid cause-effect implications. But effect size is probably too firmly embedded in the educational research usage to change now (Shaver, 1991).

size must be estimated, either the population value or the minimum result that would indicate practical significance. Once this is done, the researcher can manipulate sample size, alpha level, whether the alternative hypothesis is directional or nondirectional, and even the magnitude of the estimated effect size to obtain a desired level of power. All seem to be rather meaningless exercises, ar intellectual game, once the effect size of interest has been specified (a point to which I will return). The concern should be whether an anticipated effect size is obtained, not how to manipulate design and analysis elements so that the result, if obtained, will be statistically significant. To focus attention on the appropriate issue, S. A. Cohen and Hyman (1981) have insisted that their doctoral students specify an effect size, not just an alpha level, as the criterion against which to judge results.

How About Replicability?

Something else which a test of statistical significance is not, is an indication of the probability that a result would be obtained upon replication of the study. A test of statistical significance yields the probability of a result occurring under the null hypothesis, not the probability that the result will occur again if the study is replicated. Carver's (1978) treatment should have dealt a death blow to this fallacy, too.

With the randomization model, one has no way of knowing how close a particular result is to the population parameter. The more extreme a statistic is in the sampling distribution, the less likely it is to be reproduced upon replication. That is, with the continued drawing of random samples (and especially with continued random assignment when not conjoined with random sampling, so there is lack of control of the hypothetical populations from which the assignments are being made), there could be

considerable fluctuation in the statistics obtained. Moreover, sampling (i.e., random) error is compounded by the experimental circumstances that make it difficult in educational settings to implement a previous design with fidelity, including the valid reproduction of a treatment, across replications. These difficulties, coupled with the limited information from a randomness-based probability statement, are why statistical significance does not indicate the reliability, or replicability, of a result. Unfortunately, the contrary assumption has been common, encouraging a one-shot approach to research (Shaver, 1979).

Statistical significance not only provides no information about the probability that replications of a study would yield the same result, but is of little relevance in judging whether actual replications yield similar results. Similar probabilities could be based on quite different results and different probabilities could be based on identical results (see Rosenthal, 1991). For that reason, a recommendation that statistical significance and the statistical power of tests of statistical significance, in addition to effect sizes, be reported in replications (Rosenthal, 1991) makes little sense: The question of interest is whether an effect size of a magnitude judged to be important has been consistently obtained across valid replications. Whether any or all of the results are statistically significant is irrelevant.

The Persistence of Tests of Statistical Significance

Tests of statistical significance, then, yield little valid information pertinent to the questions of interest in most educational research.* As I

* Projects, such as the National Assessment of Educational Progress, in which random sampling techniques are used in order to estimate population parameters are notable exceptions.

noted 12 years ago (Shaver, 1980), the continued reliance on inferential statistics in educational research in the face of the many criticisms would surely make an excellent study in intellectual history and the sociology of ideas. There is, undoubtedly, a complex web of causal factors; it is an oversimplification to claim that the fault lies with journal editors who discourage the submission of reports of statistically nonsignificant results and of replications and who do not accept articles in which quantitative data are reported without tests of statistical significance.

Errors of Omission

There are, undoubtedly, many subtle factors involved in the continued use and misuse of tests of statistical significance. For example, although I have yet to see an article in which it is argued that a test of statistical significance does more than provide information on the likelihood of a result occurring under the null hypothesis, an error of omission is common. That is, there is a tendency to not remind readers of the central position of random sampling in the logic of the commonly used tests of statistical significance (e.g., J. Cohen, 1990) or to mention randomness at one point in the discussion, but ignore it elsewhere (e.g., Berk & Brewer, 1978, pp. 190-191; Carver, 1978, pp. 381-382, 385; J. Cohen, 1990, e.g., pp. 1307, 1310; S. A. Cohen & Hyman, 1980). Although the authors may understand that randomness is central to the meaningful interpretation of tests of statistical significance, many readers will miss that point when it is omitted from much of the discussion.

Statistics Courses

That educational researchers might not read "with randomization" into a statement such as, "What [statistical significance] tells us is the

probability of the data, given the truth of the null hypothesis . . ." (J. Cohen, 1990, p. 1307), is not surprising in light of the training that many receive. Statistics textbooks are sometimes the source of "myths and misconceptions" that lead researchers to overinterpret inferential statistics (Brewer, 1985). Moreover, statistics courses, like textbooks, are frequently geared almost solely to helping students recognize the types of analysis situations to which various inferential statistics are applicable and to interpreting the results in terms of statistical significance. They are not aimed at encouraging students to question the uses or usefulness of tests of statistical significance, much less to reflect on the role of random sampling in their inferential tests.

The tendency for statistical over-analysis and unthinking interpretation is being exacerbated by the growing availability, with personal computers in practically every office, of black-box statistical solutions: The researcher puts data in and receives output, often without understanding the assumptions underlying the analysis, and is encouraged in this activity by a current emphasis on data analysis with high-speed computers at the highest possible level of statistical complexity. J. Cohen's (1990) caution that "simpler is better" is refreshing advice that will likely be ignored with complex data analyses so easily accomplished. A possible cause is the education that many educational researchers receive. Their statistics courses in particular are not philosophical in orientation. They are trained to be mechanical appliers of statistical techniques, rather than educated to be thoughtful, critical users (Dar, 1987).

Graduate Committees

Graduate students' training is often reinforced by their experiences

with their graduate committees. Faculty often insist on the application of tests of statistical significance when they are not appropriate, such as in the situation where the student will collect data on a total population. As alleged by Brewer (1985, p. 246), there is also an unthinking insistence by graduate committees on the importance of results statistically significant at the .05 level. The myth persists that only results that are statistically significant at or beyond the .05 level are important, and that "beyond" means "of greater magnitude" and/or "of greater importance." Graduate students are frequently anxious that they might not obtain statistically significant results and, consequently, their research will not be acceptable to their committees. Misunderstanding of the meaning of statistical significance is perpetuated, along with misapprehension of the importance of negative (no difference) findings in the production of knowledge, a point for discussion on another day.

Being "Scientific"

Among the reasons for the insistence on tests of statistical significance is that they provide a facade of scientism in research. For many in educational research, being quantitative is equated with being scientific. Numbers, perhaps, give a sense of security and of seemingly firm results, especially when researchers confuse the numbers with reality (Shaver, 1991, p. 94). An extension is the assumption that mathematics, including inferential statistics, are essential to science—despite the fact that some scientists (for example, physicists) and many outstanding psychologists (such as Wundt, Piaget, Lewin, and Skinner) have managed very well without inferential statistics (J. Cohen, 1990, p. 1311). Unfortunately, attention is distracted from the serious consideration of educational research as science,

including the role of replication (Shaver, 1979).

Ritualized Practice

The use of statistics has been criticized as ritualistic, because it is so often unthinking practice (e.g., J. Cohen, 1990; Seeman, 1973; Shaver, 1987), and even compared to religious ritual (Salsburg, 1985). What Salsburg noted in regard to "the religion of statistics" in the medical profession applies equally well to educational research. There are the high priests, those who have the secrets to the mysteries of God, salvation, and eternal life (of statistical inference) and they are not lightly or easily challenged. The religious practitioners go to church (start up their computers) and go through their rituals (compute their tests of statistical significance) hoping for salvation (findings at the .05 level). Overall, a sense of security pervades; the practitioners have confidence in the high priests' access to the underlying mysteries, allowing them to avoid dealing with the foreboding uncertainties of the meaning of life (or of the meaning of statistical analyses).

The term "mystery" is particularly significant in this discussion because, to many educational users, tests of statistical significance really are mystical; the probability level is a magical indication of whether results are important—that is, have achieved the sacred .05 level of statistical significance. And, just as people are often raised to accept religious affiliation and practice, rather than encouraged to question underlying tenets, so do statistics courses and other graduate education experiences, as noted above, often foster the unthinking acceptance of statistical significance practices and over-interpretations rather than challenging their validity and usefulness. The acceptance of the ritual diverts attention from

important matters such as the role of replication in knowledge verification, the reporting of effect sizes to indicate the magnitude of results, and the weighing of benefits and costs in making judgments about the educational significance of results.

Social Context

Considerable social pressure also contributes to the continued unthinking use of tests of statistical significance. I can attest, as I am sure many readers of this paper can, to the subtle as well as overt pressures that have led me to have my own doctoral candidates report tests of statistical significance when they were basically meaningless additions to interpretation (a practice I have abandoned). Rejecting the use of statistics is likely to be particularly threatening—that is, it raises fears of rejection by, even humiliation at the hands of, academic peers—for those who do not feel comfortable with mathematics or with their philosophical understanding of the logic of statistics.

Such influence is present in the broader research domain as well. Glass et al. (1981, pp. 197-199) discussed why inferential statistics are not appropriate for meta-analyses in which the reviewers are analyzing data from populations or near populations of studies. They noted, however, that Tukey had "chided them for not presenting standard errors of the more important averages," despite their reasons for not doing so. He maintained that "regardless of such complications, some rudimentary inferential calculations would be informative and useful." After that revelation, Glass et al. presented a lengthy discussion of the application of inferential statistics in meta-analysis.

Another part of the social context is professors who teach inferential

statistics courses and researchers whose reputations have depended on statistically significant results. Both have psychological investments, as well as personal stakes, in the continued legitimacy of tests of statistical significance. Carver (1978) was correct: The influence of statistical significance testing will not be easily diminished because "too many have a vested interest in it" (p. 397).

Paradigmatic Effects

Breaking out of ritual, then, is not only a problem for those at the lower levels of statistical practice. As Thompson (1987, 1989) noted, paradigms, including accepted ways of thinking about research, are difficult to examine. They come to be taken for granted as natural thought and they carry normative implications for what is appropriate thinking. Paradigms dominate the thinking of the high priests as well as their followers, in science as well as in educational research, and are resistant to change in all.

As Lightman and Gingerich (1992) noted, there is in science a strong preference for "explanations that are mechanistic, logical, and calculable" (p. 694)--certainly an apt description of tests of statistical significance. Lightman and Gingerich also suggested that even in the face of anomaly, scientists--and one might assume, educational researchers--are conservative, "reluctant to change their explanatory frameworks" (p. 694). The explanations for this inertia include social and cultural factors--as noted above and as discussed by Barber (1961), such as the force of shared beliefs and commitments, the influence of and desire for professional prestige, resistance to "nonspecialists", and adherence to schools of thought--and psychological factors, such as being comfortable with the familiar and desiring to avoid the

discomfort of cognitive dissonance (Festinger, 1957).

The resistance to paradigmatic change may be so strong that facts are not even acknowledged as anomalous and thus a reason for change. Tests of statistical significance seem to fit in that category. Despite all of the critiques, the unpleasant facts about the limited applicability of tests of statistical significance seem not to be recognized as anomalous with their use; thus, the need to upset established practices is avoided.

As one example of a paradigmatic influence in educational research, Thompson (1987, p. 4) cited the unthinking use of analysis of variance, which leads individuals to think in terms of the statistical significance of differences between or among means, instead of regression techniques that focus attention on relationships. Other examples of what appear to be paradigmatic influence are not too difficult to find. I noted above the continued inappropriate application of inferential statistics in meta-analysis when the reviewer has not sampled from a population of studies, but has done an exhaustive literature search. Sophisticated discussion of the computation of standard errors and establishing confidence intervals in meta-analysis (e.g., Hedges & Olkin, 1985; Rosenthal, 1984; Wolf, 1986) and the use of those inferential statistics in reporting meta-analytic results (e.g., Schlaefli, Rest, & Thoma, 1985) seem to me to be the result of a relentless paradigmatic influence (Shaver, 1991).

Power analysis again. Another example comes from power analysis. Even those who are vocal critics of tests of statistical significance seem unable to shake loose from the nagging thought that tests of statistical significance might have more meaning than they accord to them in their criticisms. S. A. Cohen and Hyman (1981) strongly attacked the use of statistical significance

and advocated hypothesis testing based on the prior specification of a critical effect size—"the minimum value . . . that the researcher has defined as educationally, or scientifically or practically significant" (p. 60). Yet, they also suggested the continuation of power analyses and the reporting of probability levels. J. Cohen (1990) (to whom, of course, much of the emphasis on power ana'ysis can be attributed, based on his 1969 book), laid out a devastating criticism of statistical tests of the null hypothesis, but then went on to advocate power analysis. His discussion illustrates that once one has accepted the lack of information provided by tests of statistical significance, power analysis is a vacuous intellectual game.

It is difficult to argue with J. Cohen's (1990) contention that researchers should plan their research, or with his proposal that a "tentative informed judgment" be made about the population effect size under investigation. What is questionable is his recommendation that the planning include the risk in regard to a Type I error that the researcher is willing to accept and the statistical power that is desired, so that once these items have been specified, "it is a simple matter to determine the sample size you need". Moreover, if "the sample size is beyond your resources, consider the possibility of reducing your power demand or, perhaps, the effect size, or even (heaven help us) increasing your alpha level" (p. 1310). What is the purpose of power analysis and the arbitrary manipulation of criteria in order to help ensure that the researcher will obtain a desired level of probability, when statistical significance has so little meaning?*

S. A. Cohen and Hyman (1981) also recommended the same sort of

* In that context, concern over whether to state a directional or nondirectional alternative hypothesis (Pillemer, 1991) seems to me to be an equally empty exercise.

intellectual game playing--that is, after "usually" setting a critical effect size, they then "set n which is often a feasibility decision, . . . then juggle alpha and beta error risks to try to maximize power as close to 80% (beta error = .20%) setting alpha at .05, .06, .10, etc., depending on what the power tables tell us . . ." (p. 53). Why "focus on Critical ES rather than on statistical significance without giving up the latter [underlining in the original]" (p. 64)?

If effect sizes are important because statistical significance (probability) is not an adequate indicator of the magnitude of result (or much of anything else), why play the game of adjusting research specifications so that a statistically significant result can be obtained if a prespecified effect size is obtained? As Carver, 1978, pointed out: "Researchers should ignore statistical significance testing when designing research; a study with results that cannot be meaningfully interpreted without looking at the p values is a poorly designed study" (p. 394).

Another possible illustration of paradigmatic influence was mentioned earlier: Rosenthal's (1991) recommendation that, along with effect sizes, statistical significance and the statistical power of tests of statistical significance be reported in replications. That makes little sense. The question of interest is whether an effect size of a magnitude judged to be important has been consistently obtained across replications of adequate fidelity, not whether the result from a replication was statistically significant or whether the design had adequate power for a result to be statistically significant. Carver (1978) was correct: "Replicated results automatically make statistical significance unnecessary" (p. 393).

Journal Editors

Another explanation for the persistence of statistical significance testing is the influence of journal editors. It is commonly claimed (see, e.g., Kupfersmid, 1988; Neuliep, 1991) that researchers tend not to submit results that do not reach the .05 level of statistical significance often because they deem such findings to be unimportant, but also because they believe that journal reviewers and editors are unlikely to accept the results, even with accompanying effect sizes. It is also believed that journal editors have policies against publishing replication studies.

Publication is crucial to success in the academic world. Researchers shape their research, as well as the manuscripts reporting the research, according to accepted ways of thinking about analysis and interpretation and to fit their perceptions of what is publishable. To break from the mold might be courageous but, at least for the untenured faculty member with some commitment to self-interest, foolish. As gatekeepers to the publishing realm, journal editors have tremendous power. As Carver (1978) and Kupfersmid (1988) have argued, a decline in statistical significance testing is not likely to occur until journal editors take a stand against it.

What Should Editors Do?

Editors are, of course, caught up in the tangled web of influences that have been sparsely discussed above. They are not immune to the allure and emotional power of ritual and they may themselves, be somewhat in awe of the high priests, perhaps because editors, too, may not be comfortable in their understanding of or capacity to challenge tests of statistical significance. Like the rest of us, editors have difficulty separating themselves from their educational rearing and do not like to take public stances that might result

in opprobrium from their academic colleagues.

It may be unrealistic, as Carver (1978) recognized, to expect journal editors to become crusaders for an agnostic, if not atheistic, approach to tests of statistical significance. What editors can accomplish is certainly limited by the research ecology within which they work. They should, however, be familiar with the longstanding criticisms of statistical significance testing, such as cited at the beginning of this paper, and the implications for editorial policy and practice should be given serious consideration. The result should be policy that does not rely on the assumed epistemological validity of tests of statistical significance and the off-shoot, power analysis (as Thompson's, 1987, proposed model does). Following are guidelines for such a policy and for editorial practice.

Statistical Significance

In general, authors should be encouraged, even required, to minimize statistical significance in their analyses and interpretations. Reports of tests of statistical significance that are not based on randomized samples (randomly selected and/or randomly assigned) should not be published. For those studies in which randomization was an element in the design, authors should be restrained from interpretations that go beyond the legitimate conclusion in regard to whether the result is a likely occurrence assuming the null hypothesis to be true. The use of tests of statistical significance to reject the null hypothesis, to declare a particular result to be nonchance under the null hypothesis, to indicate the probability that an alternative hypothesis is true, as indicators of the probability that results are replicable, as measures of the magnitude of the result, or as indications of treatment effectiveness should not be accepted.

Effect Sizes

Even with randomness, authors should be required to state expectations for (hypotheses about) quantitative results as critical effect sizes (S. A. Cohen & Hyman, 1981). (Effect sizes other than the standardized mean difference are available [see, e.g., J. Cohen, 1988; Shaver, 1991].) Effect sizes should be required in reporting the findings from quantitative research, along with descriptive statistics (and verbal descriptions) that will help readers to interpret results and to evaluate the authors' interpretations and conclusions about the populations and settings to which the study findings might be generalized. In short, studies should be published without tests of statistical significance, but not without effect sizes. And, it should be made clear that, with effect sizes specified, power analysis is not relevant.

Equally important, authors should be required to provide justification for their interpretations of the educational significance of effect sizes. There already is a tendency to use criteria, such as J. Cohen's (1988) standards for small, medium, and large effect sizes, as mindlessly as has been the practice with the .05 criterion in statistical significance testing. Consideration of the value of the outcome--including the validity of assessment--and the human as well as financial costs in producing the outcome should be mandatory. In addition, authors should be asked to confront the reproducibility of the result as an element of its educational significance (Shaver, 1991).

Replication

The mention of reproducibility leads directly to replication, widely agreed upon as a crucial element of science but largely missing from reports of educational research (e.g., Shaver & Norton, 1980a, b). Editors should not

only actively encourage the reporting of replications, but in many instances demand replication before results can be published. In doing so, however, they should not confuse replication with reviews of research. Even if done in a meta-analytic framework, reviews are not equivalent to replications, despite implications to the contrary by some authors (e.g., Bangert-Drowns, 1986, p. 398; Carlberg & Miller, 1984, pp. 9, 10; Fiske, 1983, p. 67; Hedges & Olkin, 1985, p. 3; Hunter & Schmidt, 1990, p. 37; Jackson, 1980, p. 445; Rosnow & Rosenthal, 1989, pp. 1280-1281; J. Cohen, 1990, p. 1311). As I have noted elsewhere (Shaver, 1991), "the post hoc assembly of studies that have few, if any, planned connections and [that have] unknown differences among them is [not] an adequate substitute for purposely designed replications" (p. 91). Reviews of the literature have their place, but they do not have the explanatory power of planned replications.

A proposal for a shift in editorial policy to an emphasis on replications raises several issues. One is the matter of space: It may take more journal space to publish study descriptions that are adequate to allow replication and to describe replications sufficiently to allow the evaluation of their fidelity. Moreover, with replications given higher priority, fewer "original" studies could be published, requiring more critical judgments about study validity and importance. In the long run, the outcome should be better studies, an increased number of replication studies, and greater knowledge productivity. However, careful attention to standards will be crucial, raising other issues to be addressed by journal editors.

All studies are not worth replicating; replication will not make a trivial study worthwhile. And not all replications of potentially important studies will be worth publication. Criteria must be specified by which to

judge which initial studies are nontrivial, when a direct replication is adequately replicative, when a sufficient number of independent replications have been published to establish the reliability of a finding, and whether a systematic replication makes a substantial contribution to generalizability (e.g., Rosenthal, 1991). Kupfersmid's (1988) proposal that manuscripts first be submitted to editors without the Results and Discussion sections, so that the publication decision would be focused on study justification and design quality, is as relevant to replications as to initial studies. The logistics involved in handling acceptances and later full submissions may, however, place unrealistic demands on editors.

Type of Significance

An easily instituted editorial policy, proposed by Carver (1978, p. 395), is to insist that authors always use a modifier for the term "significant". Authors should not be allowed to make statements such as, "a confidence interval . . . tells you incidentally whether the effect is significant . . ." (J. Cohen, 1990, p. 1310). As any reader of the literature is likely to have observed, it is more common to use the term significant without a modifier than with one.

Insisting upon a modifier would make it more evident to authors that they should contemplate whether their results have educational or practical significance as well as statistical significance or, conversely, whether results lacking statistical significance might not have educational or practical significance. At the same time, readers would be alerted against overinterpreting a statistically significant result as necessarily educationally significant. Sensitivity to the important epistemological issues underlying the use of tests of statistical significance might also be

increased.

Guideline Examples

Another rather modest editorial action that might have important outcomes would be to delete misleading examples from guidelines for authors. For example, as Becker (1991) pointed out, the Publication Manual of the American Psychological Association (1983) does not provide a good model of reporting, with horrendous examples such as, "tne first grade girls reported a significantly greater liking for school" (p. 81) and "the analysis of variance indicated a significant retention interval effect" (p. 81). Moreover, the Manual discourages "the reporting of negative results" (p. 19), promoting the idea that only statistically significant results are important.

Conclusion

J. Cohen (1990, p. 1311) suggested that changes in the way people conceptualize, conduct, and report research take time, as evidenced by the over 40 years that elapsed before Gossett's t test made its way into statistics textbooks. However, criticisms of statistical significance testing have been around longer than 40 years and the practice seems to continue unabated, despite a slight increase in the reporting of effect sizes. It is probably too much to hope that 10 years hence the participants in an AERA symposium on tests of statistical significance will be able to cite evidence that the criticisms are having a noticeable impact on educational researchers' planning of analysis and reporting of quantitative results.

The reasons for the persistence of tests of statistical significance are complex. The blame for the continuation of their use cannot be laid on journal editors alone. Nevertheless, given the role of editors in determining what gets published, they could be a powerful influence for more rational use

of tests of statistical significance. Possible actions range from a rather modest insistence on the use of a modifier with the term significance to the more far-reaching policy that reports of tests of statistical significance be restricted to those based on randomized samples. Researchers would be encouraged to analyze more carefully the importance of their results and to demonstrate empirically the reliability and generalizability of study outcomes.

There are sufficient logical grounds, laid out in authoritative criticisms, on which to base changes in editorial policy to minimize, if not eliminate, the use of tests of statistical significance. No doubt action to limit the inappropriate use of tests of statistical significance would be stressful for editors, given the strength of statistical significance testing in the dominant educational research paradigm and the social and psychological pressures that encourage the continuation of the inferential ritual. Journal publication is, however, the most focused point in the process of research production and reporting, and editors have an historic opportunity to influence the course of educational and psychological research and the production of research-based knowledge in a way not available to researchers, textbook authors, or statistics professors. Modifications in publication policies have great potential for effecting the changes in the use of statistical significance that have been called for over and over but largely ignored. The question is, who will lead the way?

References

Bakan, D. (1967). On method: Toward a reconstruction of psychological investigation. San Francisco: Jossey-Bass.

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic methods. Psychological Bulletin, 99, 388-399.

Barber, B. (1961). Resistance by scientists to scientific discovery. Science, 134, 596-602.

Becker, G. (1991). Alternative methods of reporting research results. American Psychologist, 46(6), 654-655.

Berk, R. A., & Brewer, M. (1978). Feet of clay in hobnail boots: An assessment of statistical inference in applied research. Evaluation Studies Review Annual, 3, 190-214.

Bracey, G. W. (1991). Sense, non-sense, and statistics. Phi Delta Kappan, 73(4), 335.

Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? Journal of Educational Statistics, 10(3), 252-268.

Campbell, K. E., & Jackson, T. T. (1979). The role of and need for replication research in social psychology. Replications in Social Psychology, 1(1), 3-14.

Carlberg, C. G., & Miller, T. L. (1984). Introduction. The Journal of Special Education, 18, 9-10.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48(3), 378-399.

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum.

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

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

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, S. A., & Hyman, J. S. (1981, April). Testing research hypotheses with critical ES instead of statistical significance in educational research. Paper presented at the annual conference of the American Educational Research Association, Los Angeles, CA.

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

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. American Psychologist, 42(2), 145-151.

Ferguson, G. A., & Takane, Y. (1989). Statistical analysis in psychology and education (6th ed.). New York: McGraw-Hill.

Festinger, L. (1957). A theory of cognitive dissonance. Evanston, IL: Row, Peterson.

Fiske, D. W. (1983). The meta-analytic revolution in outcome research. Journal of Consulting and Clinical Psychology, 51, 65-70.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5(10), 3-8.

Glass, G. V, & Hopkins, K. D. (1984). Statistical methods in education and psychology (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Glass, G. V, McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Harcum, E. R. (1989). The highly inappropriate calibrations of statistical significance. American Psychologist, 44(6), p. 964.

Hays, W. L. (1973). Statistics for the social sciences (2nd ed.). New York: Holt, Rinehart and Winston.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. New York: Academic Press.

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.

Jackson, G. B. (1980). Methods for integrative reviews. Review of Educational Research, 50, 438-460.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43(8), 635-642.

Lightman, A., & Gingerich, O. (1992). When do anomalies begin? Science, 255, 690-695.

McHugh, R. B. (1964). Need the randomized block design be replicated? The Journal of Experimental Education, 33(2), 169-174.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. Philosophy of Science, 34, 103-115.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher, 18(2), 5-11.

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

Neuliep, J. W. (Ed.). (1991). Replication research in the social sciences. Newbury Park, CA: Sage.

Pillemer, D. B. (1991). One- versus two-tailed hypothesis tests in contemporary educational research. Educational Researcher, 20(9), 13-17.

Rosenthal, R. (1984). Meta-analytic procedures for social research. Beverly Hills, CA: Sage.

Rosenthal, R. (1991). Replication in behavioral research. In J. W. Neuliep (Ed.), Replication research in the social sciences (pp. 1-30). Newbury Park, CA: Sage.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological sciences. American Psychology, 44, 1276-1284.

Salsburg, D. S. (1985). The religion of statistics as practiced in medical journals. The American Statistician, 39(3), 220-223.

Schlaefli, A., Rest, J. R., & Thoma, S. J. (1985). Does moral education improve judgment? A meta-analysis of intervention studies using the Defining Issues Test. Review of Educational Research, 55, 319-352.

Seeman, J. (1973). On supervising student research. American Psychologist, 28, 900-906.

Shaver, J. P. (1979). The productivity of educational research and the applied-basic research distinction. Educational Researcher, 8(1), 3-9.

Shaver, J. P. (1980, April). Readdressing the role of statistical tests of significance. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Shaver, J. P. (1985a). Chance and nonsense: A conversation about interpreting tests of statistical significance, Part 1. Phi Delta Kappan, 67, 57-60.

Shaver, J. P. (1985b). Chance and nonsense: A conversation about interpreting tests of statistical significance, Part 2. Phi Delta Kappan, 67, 138-141. Erratum, 1986, 67, 624.

Shaver, J. P. (1987, October). Mandates for curriculum and instruction from educational and psychological research. Keynote address presented at the annual meeting of the Educational Research Association, Park City, UT.

Shaver, J. P. (1991). Quantitative reviewing of research. In J. P. Shaver (Ed.), Handbook of research on social studies teaching and learning (pp. 83-95). New York: Macmillan.

Shaver, J. P., & Norton, R. S. (1980a). Populations, samples, randomness, and replication in two social studies journals. Theory and Research in Social Education, 8(2), 1-20.

Shaver, J. P., & Norton, R. S. (1980b). Randomness and replication in ten years of the American Educational Research Journal. Educational Researcher, 9(1), 9-15.

Signorelli, A. (1974). Statistics: Tool or master of the psychologist? American Psychologist, 29, 774-777.

Skinner, B. F. (1956). A case history in scientific method. American Psychologist, 11, 221-223.

Thompson, B. (1987, April). The use (and misuse) of statistical significance testing: Some recommendations for improved editorial policy and practice. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Thompson, B. (1989). The place of qualitative research in contemporary social science: The importance of post-paradigmatic thought. Advances in Social Science Methodology: A Research Annual, 1, 1-42.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? American Psychologist, 24(2), 83-91.

Winch, R. F., & Campbell, D. T. (1969). Proof: No. Evidence? Yes. The significance of tests of significance. The American Sociologist, 4, 140-143.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). Statistical principles in experimental design (3rd ed.). New York: McGraw-Hill.

Wolf, F. M. (1986). Meta-analysis: Quantitative methods for research synthesis. Beverly Hills, CA: Sage.