Harvard Educational Review, Vol 48, Issue 3, Pages 378-399, 1978

The Case Against Statistical Significance Testing[1]

RONALD P CARVER
University of Missouri- Kansas City

In recent years the use of traditional statistical methods in educational research has increasingly come under attack. In this article, Ronald P Carver exposes the fantasies often entertained by researchers about the meaning of statistical significance. The author recommends abandoning all statistical significance testing and suggests other ways of evaluating research results. Carver concludes that we should return to the scientific method of examining data and replicating results rather than relying on statistical significance testing to provide equivalent information.

Statistical significance testing has involved more fantasy than fact. The emphasis on statistical significance over scientific significance in educational research represents a corrupt form of the scientific method. Educational research would be better off if it stopped testing its results for statistical significance.

The case against statistical significance testing has been developed by many critics (see Morrison & Henkel, 1970b). For example, after a detailed analysis Bakan (1966) concluded that "the test of statistical significance in psychological research may be taken as an instance of a kind of essential mindlessness in the conduct of research" (p. 436); and as early as 1963 Clark made the following comment after comparing various statistical viewpoints: "The null hypothesis of no difference has been judged to be no longer a sound or fruitful basis for statistical investigation ….. Significance tests do not provide the information that scientists need, and, furthermore, they are not the most effective method for analayzing and summarizing data" (pp. 466, 469). Shulman (1970) admonished that "the time has arrived for educational researchers to divest themselves of the yoke of statistical hypothesis testing" (p. 389), and more recently Coronach (1975) argued that "the time has arrived to exorcise the null hypothesis" (p. 124). Yet none of the above critics appears to have had an impact upon the methods of researchers in education. An informal analysis of the twenty-nine empirical articles published in the 1977 volume of the *American Educational Research Journal* shows

[1] The preparation of this paper was supported in part by the US Office of Naval Research, Contract No N00014-75-C-0958. The constructive criticisms of Martin Levit, James Hoffman, Daniel Tira, Joseph Wolff, Richard Cahoon, Ron Karraker, Bill Ghiselli, Bill Lewis, Jack Lutz, Roger Carlson, Robert Leibert, Marilyn Eanet, David Rindskopf, and students of the author are gratefully acknowledged.

that all but two of them used some test of statistical significance. Apparently the case against such testing will have to be stated more loudly and more clearly to a wider audience if it is to have any effect.

In the following sections I will discuss the definition of statistical significance testing and the fantasies often associated with its interpretation. Since not all users of statistical significance interpret their results in erroneous ways, I will also discuss the pros and cons of the more moderate use of statistical significance and argue that, ultimately, all forms of such testing should be abandoned. Finally, I will consider why statistical significance testing continues to flourish and suggest ways to govern its use in the future.

Two qualifications are in order. First, from a statistician's point of view, statistical significance testing can involve more than one procedure because it has evolved from more than one source (see the excellent review by Clark, 1963). Yet current research practices in education, as well as most introductory statistics textbooks, confine themselves almost exclusively to one procedure. Second, I do not mean to suggest that educational research is more deserving of criticism than other areas of research. This critique applies equally to all fields that use statistical significance testing in conducting research, for example, psychology, sociology, physiology, and biochemistry.

Facts about Statistical Significance

Unfortunately, only a small proportion of the specialists in education who use and interpret statistical significance testing understand what it is or how to discriminate between what it is and what it is erroneously interpreted to be. In this section, the facts about the concepts underlying statistical significance will be presented.

In the example that follows, the *t* test for mean differences will be used to make the discussion as simple as possible. The same general logic applies to other tests of statistical significance, such as the *F* test in analysis of variance and chi square. In this example, a mean has been calculated for an experimental group and another mean for a control group. these means generally come from an experiment like the one shown in Figure 1. In the box under the caption, Example Experiment, the procedures employed in research are described in abbreviated form. A sample of subjects is randomly selected from a population called the original population. These subjects are then randomly assigned to one of two groups, *X* and *Y*. Group *X* is administered the experimental treatment and is called the experimental group. Group *Y* is given no treatment and is called the control group. To make the example concrete we will use children as the original population and schooling as the treatment. The hypothetical research question is: Has the treatment, attending school, effectively increased reading ability? If the treatment has not been effective, then situation B in Figure 1 represents the situation best; group *X* is no different from group *Y*, and neither differs from the original population of children.

[Insert Figure 1]

Situation A in Figure 1 represents the situation best when the treatment is effective. Here sample *Y* still represents the original population, but sample *X* no longer does – it now represents another population called the experimental population. Of course, the children are the same even though they have had schooling, but we treat them as representing a population that has become substantially different with respect to reading ability because they have attended school.

The latter outcome is the result predicted under the research hypothesis, that is, "schooling increases reading ability". The treatment is schooling, and the effect is increased

reading ability as measured by the difference between the mean reading ability of the experimental group and the mean reading ability of the control group. From the research hypothesis it can be predicted that the two means will differ because they represent two different populations; the mean reading ability of the population of children who have attended school should be higher than the mean reading ability of those who have not.

No difference between the means of *X* and *Y* would be predicted using the null hypothesis, "schooling has no effect on reading ability". Under the null hypothesis it can be predicted that the mean of the experimental group will differ from the mean of the control group only to the extent that sampling fluctuations inevitably occur when samples are drawn from a common population.

Understanding sampling fluctuation, often called sampling error, is essential to understanding statistical significance testing. Suppose a box contains nine pieces of paper with numbers from 1 to 9 written on them. The mean of this population of numbers would be 5. If a sample size of three is drawn from the box and the mean of these three numbers is calculated, it is not likely to be 5 because of sampling fluctuation. Suppose the three numbers in the sample were 5, 7, and 9. the mean of the sample would be 7, and the difference between the sample mean and the population mean would be 2, that is, the difference between 7 and 5. This difference is a chance difference and it is called a sampling error. Because the members in a population vary, the mean of the sample is likely to differ somewhat from the mean of the population. The average sampling error becomes smaller as the size of the sample becomes larger and it also becomes smaller as the variation of the numbers in the population gets smaller.

In research, we do not know if the two sampled groups both represent the same population, but the null hypothesis says that we will *assume* that they do. The null hypothesis states that the experimental group and the control group are not different with respect to reading ability and that any difference found between their means is due to sampling fluctuation. Statistical significance testing sets up a straw man, the null hypothesis, and tries to knock him down. We hypothesize that two means represent the same population and that sampling or chance alone can explain any difference we find between the two means. On the basis of this assumption, we are able to figure out mathematically just how often differences as large or larger than the difference we found would occur as a result of chance or sampling. If differences as large or larger than the one we found occur very rarely using our straw man hypothesis, then we will reject it; sampling, or chance, is no longer considered a good explanation for the cause of the difference between the means. Rejecting this straw man hypothesis, or null hypothesis, is the same as rejecting the idea that the two groups are essentially equivalent. Researchers usually want to reject the idea that the two means are essentially equal or represent the same population. They seek support for the research hypothesis that the treatment made the experimental subjects different and that the experimental subjects now represent a changed population with respect to reading ability.

When the null hypothesis is used in research, the known variability in the sampled groups can be used to *estimate* the unknown variability in the assumed common population. Using this estimate of the population variability and the known sample size, we can mathematically calculate how often we would expect to find mean differences – sampling errors – of any particular size. The calculations from a *t* test provide a *p* value, such as $p = 0.05$; it is a number which tells us the *proportion* of the time that we can expect to find mean differences as large as or larger than the particular sized difference we get when we are sampling from the same population assumed under the null hypothesis. For example, suppose *p* was calculated to be 0.02 when there were 15 students in an experimental group with a mean of 76 and a standard deviation of 4.3, and 15 students in a control group with a mean of 72 and a standard deviation of 4.3. In this pair of samples, the difference between the means of the

two groups is 4 (76 – 72 = 4). Since *p* has been calculated to be 0.02, this means that 2 per cent of the time when sampling a pair of means from the same population we would expect to find the pair to differ by as much as 4 or more. We would expect to find the pairs of means differing by about 4, 5, 6, or larger in 2 per cent of the pairs of samples, and we would expect to find the pairs of means to differ about 3, 2, 1, or 0 in 98 per cent of the pairs of samples. A *p* value of 0.02 in this hypothetical example would mean that we would expect to find sampling errors (chance differences) as large as 4 or larger in 2 pairs of samples out of every 100 pairs of samples we took.

The *p* value, which can only be calculated by assuming the truth of the null hypothesis, is what researchers use to decide whether or not to reject the truth of the null hypothesis. The straw man hypothesis can only be knocked down by finding extremely small *p* values. In educational research, a *p* value of 0.05 (5 per cent, or 5 times out of 100) is often used as a cut-off decision point for deciding whether an event is rare or not. This cut-off is then referred to as the 0.05 level of statistical significance. When something happens 10, 18, or 33 times out of 100 occasions, for example, these frequencies would be examples of non-rare frequencies. Therefore, if *p* values are larger than 0.05, for example, 0.10, 0.18, 0.33, then differences this large or larger are not considered rare because they occur relatively frequently when sampling. Frequencies of 4, 2, or 1 out of 100 occasions, for example, would be considered rare frequencies because they occur less than 5 times in 100 occasions. In our example ($p = 0.02$), we would rarely expect to find pairs of means to differ as much as 4 when sampling under the conditions of the null hypothesis ($p \leq 0.05$), so we conclude that it is not reasonable to expect that the null hypothesis is true. Therefore, we reject it.

When an experimenter finds a low *p* value and rejects the null hypothesis, he or she knows that a calculated risk is being taken. When the null hypothesis is true and therefore both samples do represent a common population, the experimenter will make a mistake and reject the null hypothesis 5 percent of the time (that is, at the 0.05 level of statistical significance). Assuming we know for sure that the straw man should remain standing, our procedures will be right 95 per cent of the time when the null hypothesis is really true. Thus, the experimenter knows that the odds are 95 to 5 against making a wrong decision when the null hypothesis is really true. Yet there is no way in practice that we can be absolutely sure the null hypothesis is true. If we could be sure, we would never test for statistical significance at all.

Statistical significance simple means statistical rareness. Results are "significant" from a statistical point of view because they occur very rarely in random sampling under the conditions of the null hypothesis. A statistically significant mean difference between two research groups at the 0.05 level indicates the following: if we assume that the two research groups are random samples representing the same hypothetical population which has properties that can be estimated from properties of the groups themselves, and if we assume that we sampled 100 sets of two groups from this same hypothetical population, then we would expect to find the mean difference between the two research groups to be larger than 95 of the 100 sampled from the hypothetical population. A statistically significant result means that the probability is low that we would get the type of research we got, given that the null hypothesis is true.

Fantasies about Statistical Significance

In the preceding section we dealt with the meaning of statistical significance. By itself, statistical significance means little or nothing. It becomes important only when it is used to make inferences. We will now discuss three fanciful inferences about the causal role of chance, the replicability of results, and the validity of a research hypothesis.

Odds-Against-Chance Fantasy

The first of the three fantasies can be called the "odds-against-chance" fantasy. It is an interpretation of the *p* value as the probability that the research results were due to chance, or caused by chance (see Wilson, 1961). As has been explained, the *p* value is the probability of getting the research results when it is first assumed that it is actually true that chance caused the results. It is therefore impossible for the *p* value to be the probability that chance caused the mean difference between two research groups since (a) the *p* value was calculated by assuming that the probability was 1.00 that chance did cause the mean difference, and (b) the *p* value is used to decide whether to accept or reject the idea that probability is 1.00 that chance caused the mean difference.

A *p* value of 0.05 actually means that the odds are 1 in 20 of getting a mean difference this large or larger, and the odds are 19 in 20 of getting a mean difference this large or smaller *if* the two samples represent the same population, as in situation B. We do not know how to estimate what the odds are that situation B is true. We do not know what the odds are that the result can be attributed to chance. Such an interpretation would have to be derived from a hypothetical situation where the 100 pairs of means were divided into two groups where the null hypothesis was true for one group of means and not true for the other. Then if we could estimate the number of mean differences that occurred by chance, for instance, if 5 resulted when the null hypothesis was true, we would automatically have an estimate of those that did not occur by chance, that is, 95 resulted when the null hypothesis was not true. However, we never in fact have this latter hypothetical situation. When calculating the *p* values, we assume situation B to be true in all 100 cases; in other words, we assume that chance or sampling is responsible for all 100 of the mean differences. We assume that all the odds are in favor of chance causing the results.

Generations of students have been misdirected by textbook writers into the above erroneous interpretation of statistical significance. For example, most educational researchers have taken a principles-of-testing course, and one of the favored textbooks for this course (Anastasi, 1976) is in its fourth edition. Consider the following quotation from that text:

> To say that the difference between two means is significant at the 0.01 level indicates that we can conclude, with only one chance out of 100 of being wrong, that a difference in the obtained direction would be found if we tested the whole population from which our samples were drawn. (p. 109)

This is one way of stating the odds-against-chance fantasy. Hebb (1966) has expressed the odds-against-chance fantasy more explicitly: "When one encounters the statement that a difference is significant, it signifies, by convention, that the probability is at least 19 to 1 against this being due to the operations of chance in obtaining our sample" (p. 173). The 19-to-1 odds come from the 0.05 level, that is, 0.95 and 0.05 divided by 0.05 are 19 to 1. A statistically significant result at the 0.05 level is conventionally interpreted to mean that in only 5 times out of 100 this result would be due to chance, or sampling, but this is a completely erroneous conclusion.

Cronbach and Snow (1977) succinctly summarize this particular fantasy in terms of probability statements:

> A *p* value reached by classical methods is not a summary of the data. Nor does the *p* value attached to a result tell how strong or dependable the particular result is ….. Writers and readers are all too likely to read 0.05 as p(H|E), "the probability that the Hypothesis is true, given the Evidence." As textbooks on statistics reiterate almost in

> vain, *p* is p(E|H), the probability that this Evidence would arise if the [null] hypothesis is true. Only Bayesian statistics yield statements about p(H|E). (p. 52)

This is perhaps the most important and least understood principle of statistical significance testing. The following example is an attempt to make this principle clearer.

What is the probability of obtaining a dead person (label this part D) given that the person was hanged (label this part H); this is, in symbol form, what is p(D|H)? Obviously, it will be very high, perhaps 0.97 or higher. Now, let us reverse the question. What is the probability that a person has been hanged (H), given that the person is dead (D); that is, what is p(H|D)? This time the probability will undoubtedly be very low, perhaps 0.01 or lower. No one would be likely to make the mistake of substituting the first estimate (0.97) for the second (0.01); that is, to accept 0.97 as the probability that a person has been hanged given that the person is dead. Even though this seems to be an unlikely mistake, it is exactly the kind of mistake that is made with interpretations of statistical significance testing --- by analogy, calculated estimates of p(H|D) are interpreted as if they were estimates of p(D|H), when they are clearly not the same.

In statistical significance testing, we ask: what is the probability of obtaining a large mean difference (label this D') between two samples, if the two samples were obtained from the same population (label this H_o, the usual symbol for the null hypothesis); that is, what is $p(D'|H_o)$? In concrete terms, let us say that the *p* value obtained from a statistical significance test was 0.05 so that $p(D'|H_o) = 0.05$. Now if we reverse the question to ask what is the probability that two obtained groups were sampled from the same population, we have the question that most people want to answer and assume they have answered when they calculate the *p* value from statistical significance testing. In essence they are asking what the probability is that the null hypothesis, H_o, is true, given the type of large mean difference we have obtained, or, what is $p(H_o|D')$? The *p* value that was obtained from statistical significance testing, for example, $p(D'|H_o) = 0.05$, is used as an answer to the reverse question as well. This is a fantasy, however, because the *p* value that results from statistical significance testing is $p(D'|H_o)$, not $p(H_o|D')$.

Replicability or Reliability Fantasy

The replicability or reliability fantasy is even more radical than the odds-against-chance fantasy. It is an interpretation of statistical significance as the probability of obtaining the same results whenever a given experiment is replicated (Bakan noted this error in 1966, and Lykken elaborated upon it in 1968).

An example of this fantasy comes from a recent introductory statistics textbook by Nunnally (1975):

> If the statistical significance is at the 0.05 level, it is more informative to talk about the *statistical confidence* as being at the 0.95 level. This means that the investigator can be confident with odds of 95 out of 100 that the observed difference will hold up in future investigations. (p. 195)

In essence, this explanation says that the complement of the *p* value, that is $1 - p$ ($1 - 0.05 = 0.95$), yields the probability that the results are replicable or are reliable. According to Nunnally, the probability that a certain mean difference can be replicated (R) – given the obtained mean difference – is $1 - p$; in the above example, Nunnally would say that $p(R|D') = 0.95$, but this is wrong! Too often, researchers will get a statistically significant difference and will refer to this difference as a "reliable difference" or say that "the results were reliable"

(Gold, 1969). Again, the *p* value resulting from statistical significance testing is $p(D'|H_o)$, not $p(R|D')$, which is the probability that the results are replicable or reliable.

Whether or not the results are replicable or reliable actually depends upon whether the important variables can be controlled and manipulated in exactly the same way in order to give the same results. Certainly, there is no magic in the numbers collected and analyzed using the assumptions of the null hypothesis that allows us to infer anything about the probability that another researcher will be able to get a similar mean difference. A researcher who gets a statistically significant result in favor of a particular teaching technique may find that another researcher cannot replicate the result simply because the result was attributable to the skill of the instructor working with the experimental group and not to the teaching technique itself. Too often statistical significance is substituted for actual replicative evidence; too often statistical significance covers up an inferior research design.

Nothing in the logic of statistics allows a statistically significant result to be interpreted as directly reflecting the probability that the result can be replicated. It is a fantasy to hold that statistical significance reflects the degree of confidence in the replicability or reliability of results.

Valid Research Hypothesis Fantasy

The third and most serious fantasy is the belief that statistical significance directly reflects the probability that the research hypothesis is true. For example, a *p* value of 0.05 is interpreted to mean that its complement, 0.95, is the probability that the research hypothesis is true (Bolles discussed this erroneous interpretation in 1962). Since H_1 is commonly used as a symbol for the research hypothesis, this misapprehension can be expressed as interpreting $1 - p(D'|H_o)$ as if it were as if it were $p(H_1|D')$.

It is difficult to find examples to illustrate this fantasy because it is usually not explicit. Yet an analysis of some articles in the research literature has suggested that their publication was based primarily on a report of statistically significant results rather than on data in support of rationally defensible hypotheses (see Carver, 1976; Lykken, 1968). The *p* value resulting from statistical significance testing has been used in some studies to indicate that the research hypothesis was probably true even though the research itself was questionable in both theory and methodology.

Those researchers who succumb to the valid research hypothesis fantasy are also likely to interpret the size of the *p* value as reflecting the degree of validity of the research hypothesis, that is, the lower the *p* value such as $p \leq 0.001$, the more highly significant or valid the research hypothesis. This techniques of using the size of the *p* value as a "measure" of the significance of the results has been criticized by Bakan (1966). This fantasy is perpetuated by the common practice of referring to "significant" or "highly significant" results rather than "statistically significant" results. Too many people in education do not seem to realize that a statistically significant result at the 0.05 level has nothing directly to do with inferences about the research hypothesis. Even if the null hypothesis can be rejected, several other alternative or rival hypotheses still must be ruled out before the validity of the research hypothesis is confirmed. Only after rigorous theorizing, careful design of experiments, and multiple replications of the findings in varied situations should one contend that the probability is high that the research hypothesis is true.

Summary

The misinterpretation of, or fantasies about, statistical significance at the 0.05 level fall into three categories: (a) the probability is 0.05 that the results are due to chance, or the probability is 0.95 that the results were not caused by chance; (b) the probability is 0.95 that the results will replicate, or we can be 95 per cent confident that the results are reliable; and

(c) the probability is 0.95 that the research hypothesis is true, or we can be 95 per cent confident that our results are valid.

Properly interpreted, statistical significance testing provides a *p* value or the probability of obtaining mean differences of given sizes under the null hypothesis. Thus, the *p* value may be used to make a decision about accepting or rejecting the idea that chance caused the results. This is what statistical significance testing is – nothing more, nothing less.

Statistical Significance Testing: The Moderate and Conservative Positions

The fantasies held by radical proponents of statistical significance testing, which were presented in the preceding section, do not represent the practices of all researchers. In this section, the pros and cons of the more moderate and conservative uses of statistical significance testing will be given.

The moderate proponents of statistical significance testing contend that a researcher ought to be able to design and execute an experiment in which differences as large as were found would only happen *rarely* under chance or sampling conditions. Most people, including scientists, are more likely to be convinced by phenomena that cannot readily be explained by a chance hypothesis. It is contended that if a study is well designed and well executed, then demonstrating statistical significance should present no problem. According to this position, rejecting the null hypothesis by getting statistically significant results is ordinarily required before the experimenter can claim support for the research hypothesis; therefore rejecting the null hypothesis by obtaining statistical significance is a preliminary hurdle of the experimental process. The research hypothesis and the null hypothesis are usually considered to be mutually exclusive. If the null hypothesis can be rejected, empirical support can automatically be claimed for the research hypothesis. If the null hypothesis cannot be rejected, the research hypothesis receives no support.

One problem with the moderate position is that it allows the size of the sample to determine whether the results will be "significant" or not. A related problem, more serious in consequence, is that the moderate position involves a corrupt form of the scientific method. Some of the more conservative proponents seem to have avoided the two above problems when they interpret statistical significance but their advocacy has disadvantages as well. Each of the problems associated with the moderate and conservative positions will now be discussed.

Statistical significance ordinarily depends upon how many subjects are used in the research. The more subjects the researcher uses, the more likely the researcher will be to get statistically significant results. Nunnally (1960) stated "if the null hypothesis is not rejected, it is usually because N is too small" (p. 643); this is because it is unlikely that two groups represent *exactly* the same population with respect to the variable being measured (Bakan, 1966). Hays (1963) has amplified the point as follows: "Virtually any study can be made to show significant results if one uses enough subjects regardless of how nonsensical the content may be" (p. 326). Since the experimenter often has complete control over the number of subjects sampled, one of the most important variables affecting the results of the research, the subjective judgment of the experimenter in choosing sample size, is usually not controlled. Controlling experimenter bias is a much discussed problem, but not enough is said about the experimenter's ability to increase the odds of getting statistically significant results simply by increasing the number of subjects in an experiment. A concrete example can illustrate the importance of sample size. Assume that the research hypothesis is true, that the experimental group actually represents a population with a mean score of 76 and a standard deviation of 4.3, and that the control group actually represents a different population with a mean score of

72, and standard deviation of 4.3. It can be determined that if the samples perfectly represent their respective populations, a sample size of at least 11 is needed for statistical significance. If a researcher chooses a sample size of 10 or less, a statistically significant result would not be obtained, although a larger sample size would yield such results. In this way, the experimenter can directly control the probability of obtaining statistically significant results simply by controlling the sample size.

Because statistical significance in research ordinarily depends on sample size, trivial results are often interpreted as "significant" or important when they are simply results that would rarely happen when randomly sampling from the same population using large sample sizes. A mean difference that is small and not significant from a research standpoint can be *statistically* significant just because enough subjects were used in the experiment to make the result statistically rare under the null hypothesis. For example, the difference between one group that was instructed using an innovative curriculum and another group instructed with a regular curriculum might have been statistically significant at the 0.05 level. But there may have been 500 subjects in each group, and the difference in terms of grade equivalents on an achievement test might have been 1/200 of a grade level, a result that can easily be considered trivial because little scientific or practical importance can be attached to 1/200 of a grade level. In practice it can be regarded as a zero difference. But this kind of result is sometimes interpreted by the moderate proponents of statistical significance as being scientifically significant or important simply because it is statistically significant.

Conversely, statistically nonsignificant results are conventionally interpreted as providing no support for the research hypothesis even when the actual results support it. When a small sample is used, large differences in the results can more often occur by chance and therefore provide no statistically significant evidence in support of the research hypothesis. In the previous example, suppose there were only 10 subjects in each group and the mean difference in grade equivalents was 0.9. A difference this large that occurred over a semester of instruction probably is a significant or important difference supporting the researcher's hypothesis, even though it may not be statistically significant at the 0.05 level. At a minimum, it is important enough to justify the replication of the study to find out if approximately the same 0.9 grade level difference could be found again.

In 1931 Ralph Tyler pointed out that a statistically significant difference is not necessarily an important difference, and a difference that is not statistically significant may be an important difference. Unfortunately, we are still making the same point more than forty-five years later, and the warnings have been repeated many times since then. Gold (1969) wrote that statistically significant results do not necessarily indicate substantively important results: "Statistical significance is only a necessary but not sufficient criterion of importance" (p. 41). Here, Gold is actually trying to jump two hurdles: evaluate statistical significance first, then evaluate scientific significance. This strategy of giving the null hypothesis top priority still has the built-in danger of permitting results to be interpreted as not *scientifically* significant simply because they are not *statistically* significant. (For a more detailed refutation of this double-hurdle method set forth by Gold, see Morrison & Henkel, 1969.)

[Insert Figure 2]

Figure 2 shows schematically the arguments that have been covered up to this point. Under the caption "Scientific Method" is a flow chart that indicates how a research hypothesis is tested by collecting data and then comparing the results with those predicted from the research hypothesis. In our example, if the difference between the mean of the experimental group and the mean of the control group is in accordance with what was predicted by the research hypothesis, then this constitutes evidence in favor of the research

hypothesis. Alternative hypotheses, such as the chance hypothesis, are also considered. In the scientific method, data provide empirical evidence that allows up or down adjustments in the probability that the research hypothesis is true, p ($H_1|D'$). The theorem of Bayes is an example of a statistical procedure that, unlike statistical significance testing, deals directly with the probability that the research hypothesis is true (Bakan, 1966; Cronbach & Snow, 1977).

Rozeboom (1960) pointed out that evidence in accordance with a research hypothesis should increase the scientist's degree of belief in that hypothesis, whether the evidence is statistically significant or not. By saying this, he becomes one of the first to point out the unsuitability of the method of inference that underlies what I will call the corrupt scientific method, depicted in Figure 2. This procedure begins as the scientific method does but deviates radically at the point where data are interpreted with respect to the research hypothesis. At this point the statistical null hypothesis is interjected. If the data are not statistically significant, the null hypothesis is accepted as the most reasonable explanation for the results. The data are not even interpreted with respect to the research hypothesis. A result that is not statistically significant is automatically interpreted as providing no support for the research hypothesis, no matter how much the data tend to confirm it. If the data are statistically significant, they are considered as supporting the research hypothesis. Alternative hypotheses are also considered, as in the regular scientific method, but the chance hypothesis or null hypothesis is no longer seriously considered as an alternative hypothesis once statistical significance has been found. Future research is counted upon to rule out alternative hypotheses other than chance.

The moderate proponents of statistical significance testing could considerably strengthen their position if they would simply put the research hypothesis ahead of the null hypothesis, which would keep the focus on the relative size of the means themselves. Then the results of statistical significance testing would not distract from the ultimate basis for making inferences – the data themselves (Tukey, 1962). The statistical significance of the difference would be mentioned secondarily. Ordinarily, however, most moderates corrupt the scientific method in a way that often circumvents the ordinary benefits derived from it, although they tend to consider their practice as being one of the highest forms of science.

Another group of researchers uses statistical significance testing in a more limited way. They were described earlier as conservatives. They believe that such testing helps eliminate only one threat to the validity of research results, sampling error; and because there are many other threats, the importance of statistical significance testing is reduced. Winch and Campbell (1969), for example, list fifteen threats to internal and external validity and note that statistical significance is only relevant to one of the threats to internal validity[2]. Calling the threat of sampling error "instability", they further argue that just because instability is only one of fifteen threats to validity "certainly does not lead to the conclusion that such tests are pointless" (p. 140). The concluding paragraph in their article summarizes the conservative position:

> Some critics of tests of significance seem to be saying that since these tests do not dispose of all rival hypotheses, they are useless and misleading and should be abandoned. We reason that it is very important to have a formal and nonsubjective way of deciding whether a given set of data shows haphazard or systematic variation. If it is haphazard, there is no reason to engage in further analysis and to be concerned

[2] Campbell and Stanley (1966) distinguish between internal and external threats to the validity of experimental results. Internal validity refers to whether the difference found between the means was actually caused by the experimental treatment. External validity refers to whether the treatment used will have the same effect in other situations beyond the experiment that was just conducted.

> about other threats to validity; if it is systematic our outline shows that the analysis is not concluded with the test of significance but is just getting under way. And we believe it is important not to leave the determination of what is a systematic or haphazard arrangement of data to the intuition of the investigator. (p. 143)

The argument of Winch and Campbell in favor of using statistical significance testing does not appear to be compelling. It would be nice to have a "formal and nonsubjective way of deciding whether a given set of data shows haphazard or systematic variation" (p. 143). If, however, statistical significance testing leads us to decisions with questionable scientific validity, then formality and objectivity become meaningless. The subsequent argument about haphazard versus systematic variation is simply another version of the corrupt scientific method discussed earlier. Like Gold (1969), Winch and Campbell are arguing for a double hurdle, with the statistical significance test coming before further consideration of the results. It is better, they believe, to depend on the objectivity of statistical significance to make judgments about haphazardness than to leave such decisions up to the intuition of the investigator.

The conservative position of Winch and Campbell seems to suggest that science would profit from adopting the corrupt scientific method as a way of avoiding subjectivity. Yet it seems inconsistent to say that experimenters can be trusted with the hundreds of decisions that must be made in the design, execution, and data analysis of an experiment but that they cannot be trusted when it comes to deciding whether or not the data support the research hypothesis. Furthermore, with this position the experimenter must posit that the mean difference is zero (the null hypothesis); however, as Bakan (1966) has pointed out, several a priori reasons exist for believing that the null hypothesis is generally false. Therefore, why be diverted by this questionable hypothesis of no mean difference when a more reasonable hypothesis would be that there is a difference? But hypothesizing a difference in favor of the experimental group we are more likely to focus on the size of the mean difference to decide whether the hypothesis has support.

A modicum of justification for using statistical significance testing might be found if the threat to validity it posed were not trivial. It would certainly make things easier if educational researchers could theorize, design experiments, and collect data so well that the only important threat left to the validity of the result was instability. Unfortunately, theories are usually vaguely formulated, experimental designs are clearly vulnerable to the influence of uncontrolled variables, data-collection procedures are often situation-specific, and subjects are often erratic. These exigencies require that the instability threat to validity take its rightful place along with the other threats to internal and external validity, rather than being sent to the head of the line in terms of space and attention devoted to statistical tests. Even if we were justified in paying so much attention to the instability threat, statistical significance testing would still be of questionable use in guarding against instability. It still gives us an estimate of $p\ (D'|H_o)$ when what we want is $p\ (H_o|D')$, $p\ (R|D')$, and $p\ (H_1|D')$.

In light of the preceding problems with the moderate and conservative positions, is there any reason to do statistical significance testing? If we can control statistical significance simply by changing sample size, if statistical significance is not equivalent to scientific significance, if statistical significance testing corrupts the scientific method, and if it has only questionable relevance to one out of fifteen threats to research validity, then I believe we should eliminate statistical significance testing in our research. Such testing is not only useless, it is also harmful because it is interpreted to mean something it is not.

The Reasons Why Statistical Significance Testing Flourishes

If statistical significance is really trivial significance, why does it continue to flourish? Why has it been allowed to drive out the traditional scientific method just in those areas of science that are trying so hard to become respectable? One plausible explanation seems to be that small-sample-size research in education and psychology is on the defensive. Whenever someone tries to discredit the findings by pointing to the small size of the sample, the researcher defiantly responds that the results were statistically significant.

Another reason for the popularity of statistical significance testing is probably that complicated mathematical procedures lend an air of scientific objectivity to conclusions. Sophistication in educational research is almost synonymous with complex tests of statistical significance, such as the *F* test in an analysis of variance, Bakan (1966) has warned against accepting the myth that because such testing is mathematical, it is therefore precise and valid.

The two most influential reasons for the tenacity of statistical significance testing, however, involve replicability and the importance of differences. Statistical significance is generally interpreted as having some relationship to replication (see, for example, Coleman, 1964; Melton, 1962), and replication is the cornerstone of science (Bauernfeind, 1968; Smith, 1970). If results are due to chance, then results will not replicate. The only valid reason for considering statistical significance is to try to determine whether research results are simply a product of chance and will therefore not be replicable. Yet it is not logical to deduce that if the results are statistically significant, they will replicate, or that if the results are not statistically significant, they will not replicate. But if researchers do obtain the same result more than once, it is more reasonable to conclude that the results are not due to chance.

Since one of the primary reasons for being concerned with statistical significance is that chance is a threat to replication, replicated results automatically make statistical significance unnecessary (Bauernfeind, 1968). Stevens (1971) stated the relationship between statistical significance testing and replication this way:

> In the long run scientists tend to believe only those results that they can reproduce. There appears to be no better option than to await the outcome of replications. It is probably fair to say that statistical tests of significance, as they are so often miscalled, have never convinced a scientist of anything. (p. 440)

A researcher who wants to abandon statistical significance, but is still worried about chance and replicability, can thus solve both problems by searching for replicated evidence. The popularity of statistical significance would probable decline appreciably if it were more widely recognized that it is not a predictor of the replicability of research data. It seems best to rely upon direct evidence of replication rather than upon the myth that somehow statistical significance predicts replicability.

Statistical significance testing has also flourished because it is used to determine whether the size of a difference is important or not. Researchers often find it difficult to decide whether the differences found are large enough to be considered important. Stevens (1968), for example, asked: "Can no one recognize a decisive result without a significance test?" and added that making a "scientific decision by statistical calculation" was an "illusion of objectivity" (p. 853). Statistical significance testing has purportedly provided an objective, although inappropriate, solution to the problem of deciding whether a result is important.

A major problem involved in adjudicating the scientific significance of differences is that we often deal with units of measurement we do not know how to interpret. For example, suppose we administer a teaching effectiveness scale to two groups of teachers – one group having just received intensive teacher training, the other group receiving no such training. If the mean difference between the groups were ten points on the scale, we would ordinarily use a test of statistical significance to determine whether this difference is "significant" or

important. Seldom have we collected data about the measurement scale that would allow us to determine whether the raw-score difference of ten points is scientifically significant or not. Most scales used in psychology and education have been psychometrically developed to discriminate between individuals; they are seldom developed to measure group differences, individual change, or treatment effects (see Carver, 1974). It is practically impossible to determine whether a difference obtained on an educational measurement device is significant in the sense of being important. One way to solve the problem is to test the measure on extreme groups. For example, the teacher-effectiveness scale could be given to a group of master teachers and a group of student teachers. Comparing the mean difference between the extreme groups with the mean difference between the experimental and control groups provides a context for interpreting the treatment effect.

Because researchers often are not able to determine whether a difference is significant or not, they are too willing to let statistics provide an object and automatic solution, even though it is inappropriate. If more educational research understood that a "significant" difference is not necessarily more replicable, or more important than a statistically insignificant difference, fewer tests of statistical significance would be conducted.

Recommendations

Proponents of statistical significance would do well to abandon it, at least to the extend of abandoning the corrupt scientific method. If they must do statistical significance testing, they should do it after the results have been interpreted with respect to the research hypothesis, as the scientific method requires. There are always rival hypotheses to consider and the null hypothesis should rightfully take its place among these secondarily important hypotheses, in terms of attention and space in a research report.

At a minimum, the research hypothesis ordinarily predicts the direction of the mean difference, and the data can initially be interpreted with respect to the prediction. Better yet, the size of the effect could be measured and evaluated using absolute differences, omega squared, eta squared (see Hays, 1963), or d (see Cohen, 1977). Whenever a p value is reported, an accompanying statistic should reflect the size of the effect or the strength of the association between X and Y (see Hays, 1963). This value should be interpreted with respect to the research hypothesis regardless of its statistical significance. In addition, proponents of statistical significance testing should calculate estimates of the power of their statistical tests (see Chase & Tucker, 1976; Cohen, 1977); the power of a statistical test of the null hypothesis is the probability that a false null hypothesis will be rejected.

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. Certain superb designs, such as the time series designs discussed by Campbell and Stanley (1966), are seldom used – probably because no straightforward or neat test of statistical significance is associated with them. The best evidence relevant to the purpose of the research should be collected. The lack of an apparent, neat way of testing for statistical significance in a planned design does not mean that the research cannot be done using perfectly sound scientific techniques. Plenty of descriptive statistics are available for analysing research results with respect to research hypotheses. Some people avoid classroom research, for example, because the use of intact groups usually violates fundamental assumptions of analyses of variance; others go ahead and calculate p values from these analyses of variance even when they are completely erroneous. It would be better to disregard statistical significance in these situations where the classroom is the unit of analysis and where there are too few classrooms to get statistical significance.

Given that statistical significance testing usually involves a corrupt form of the scientific method and, at best, is of trivial scientific importance, journal editors should not require it as a necessary part of a publishable research article. On the contrary, editors should consider rejecting articles that contain this trivial information, just as they presently reject articles that contain raw data. Manuscript referees should continue to look for evidence of internal and external validity but should not allow statistical significance to be interpreted as crucial evidence supporting the stability, reliability, replicability, or importance of the results. "The stranglehold that conventional null hypothesis significance testing has on publication standards must be broken" (Rozeboom, 1960, p. 430).

The publication bias in favor of statistical significance was initially articulated by Melton (1962), when he was editor of the *Journal of Experimental Psychology*. He stated that it had been his policy as journal editor to discriminate against articles containing results that did not reach a *p* value of .01. Morrison and Henkel (1969) pointed out that the only time a significance level can be rationally decided upon is in a situation where the cost of a wrong decision can be calculated; they then made the following statement that is relevant to Melton's pronouncement:

> To insist upon the .05 or the .01 level is, then, to talk about the science of business, not the business of science. To say we want to be conservative, to guard against accepting more than 5 percent of our false alternative hypotheses as true (by rejecting less than 5 percent of our true null hypotheses) is nonsense in scientific research. (p. 137)

Journal editors and referees should never allow the author of a publication to state, as is now the current practice, that the results were "significant" when what is meant is "statistically significant" (Morrison & Henkel, 1969). The most recent edition of the American Psychological Association's publication manual (1974), used by almost all educational research journals, contains numerous misleading examples of phrasing such as "significantly greater" for "*statistically* significantly greater" and "significant retention" for "*statistically* significant retention" (p. 39). Kish (1959) recommended that the phrase "test of significance" be replaced by the proper phrase "test against the null hypothesis" (abbreviated TANH). Once the profession becomes aware of these pitfalls and begins, for example, to notice that "statistically" is missing from in front of "significant", readers will realize that such articles usually include no measure of, or interpretation of, effect size, as Katzer and Sodt (1973) recommended.

Journal editors should make sure that very little space is allotted to statistical significance testing in giving results. Despite the arguments advanced by Cronbach and Snow (1977) for always presenting the basic descriptive statistics, examples can be found in the literature where the analysis of variance summary tables are presented but the most important data (means and variances) are not reported. Results sections are often chock full of *Fş ts*, and *p* values that usually cannot be generalized beyond the experiment itself because other researchers will not use exactly the same number of subjects; therefore, the information is irrelevant to a researcher trying to find out if the data are replicable.

There is another problems involved in restricting articles to statistically significant results. Subsequent investigators who gather evidence relevant to the same research hypothesis and who find zero differences or statistically nonsignificant results cannot get their research published, and this effectively suppresses nonreplicability and other evidence invalidating the published results. Bakan (1966) treats this problem admirably and in detail.

A cursory sampling of doctoral dissertations in education reveals that this is where the corrupt scientific method is most prevalent, with the student often forced to state *all* null

hypotheses and *none* of the research hypotheses. Such practices are almost never mirrored in published research and are inappropriate for training. In 1970 Coats wrote, "Most graduate schools of education still require students to take what may be one of the most irrelevant learning experiences of their entire educational career. The requirement is the study of inferential statistics ….. inferential statistical models are inappropriate as typically used for analyzing data in educational research" (p. 6). The complete abandonment of statistical significance testing in the training of doctoral students in educational research should be seriously considered.

Researchers who advocate eliminating statistical significance testing must find ways of collecting and analyzing data that will provide convincing evidence. Some proponents of statistical significance testing would disregard as chance ten separate studies that replicated a certain mean difference, if each did not show the difference to be statistically significant. Such an extreme bias is difficult to overcome; but fortunately, most researchers, whatever their bias, are sensitive to replicative or collaborative forms of evidence and can recognize a good research study whether or not it uses statistical significance testing. As for evaluating the reliability or accuracy of the results, such statistics as standard errors and confidence intervals[3] should not be ignored. It is also desirable to build replication into the design, even though replication is a complex process (see Carver, in press; Lykken, 1968; Sidman, 1960).

When researchers abandon statistical significance testing, they will at first have a difficult time filling the vacuum and will want to know what to do instead. Morrison and Henkel (1970a), in their excellent book of readings about the controversy surrounding statistical significance testing, summarized this problem:

> What we do without the tests, then, has always in some measure been done in behavioral science and needs only to be done more and better: the application of imagination, common sense, informed judgment, and the appropriate remaining research methods to achieve the scope, form, process, and purpose of scientific inference. (p. 311)

Instead of the decisive "accept" or "reject" statements that come from the statistical significance testing, the most we should expect from our research is empirical data, such as descriptive statistics, that allow us to decide what kind of evidence we have for our research hypothesis: strong support, support, weak support, no support or disconfirming evidence.

The Future of Statistical Significance Testing

The reader may still wonder whether a compromise might not be to interpret empirical results with respect to the research hypothesis first and then to report the statistical significance of the results, thereby avoiding many of the pitfalls of the corrupt scientific method. In theory this would allow one to reap whatever small benefits there might be to using statistical significance testing. This position has superficial advantages, but they are outweighed by the disadvantages. Reporting *p* values is likely to influence the reader by erroneously suggesting that the results are significant and replicable. Students will still be forced to devote hundreds of hours to learning statistical significance testing procedures, often at the expense of learning about replication designs, confidence intervals, correlation rations, intraclass correlations, and other effect-size measures. Statistical significance testing is also likely to continue to

[3] Confidence intervals are sometimes confused with level of statistical significance (Chandler, 1957); for example, the 95 percent confidence interval might be confused with the .05 level of statistical significance when statistical significance is erroneously interpreted as signifying 95 per cent confidence, as Nunnally (1975) did.

encourage researchers to investigate hypotheses that are readily tested using research designs that permit neat statistical tests, whether the hypotheses are the most important or not. If one could no longer use statistical significance to determine the "significance" of a difference, researchers would be forced to use designs that more clearly reveal the scientific importance of a difference. Without statistical significance, researchers will be forced to grapple with the problems of scientific inference instead of those associated with statistical significance testing.

It is doubtful that much can be done in the next decade to eliminate statistical significance testing from educational research. Too much of the curriculum is devoted to it, and too many have a vested interest in it. Much of the influence in educational and psychological research comes from methodology (Bereiter, 1964), and most of the methodologists are highly sophisticated in statistical significance testing techniques. Researchers will continue to be convinced that if they do not find statistical significance they will be helpless if someone tries to discredit their research. The use of statistical tests of significance are not likely to decline until one or more journal editors speak against statistical significance testing in a manner similar to Melton (1962), who spoke in favor of it. Probably the most that can be hoped for is the abandonment of the corrupt scientific method, the addition of "statistically" in front of "significant", and the addition of effect-size measure to introductory statistics textbooks at the expense of space devoted to statistical significance testing.

In conclusion, a statistically significant result can be a trivial result; it should never be referred to as a "significant" result. A statistically significant result erroneously suggests that the probability is low that the result was caused by chance, and that the probability is high that the result is reliable, valid, and important. Even without the fanciful interpretations, statistical significance testing usually involves a serious breach of the scientific method; therefore it is more accurate to say that statistical significance testing uses a corrupt form of the scientific method. Even if properly used in the scientific method, educational research would still be better off without statistical significance testing. Case closed – for now.

References

American Psychological Association. *Publication Manual* (2nd ed). Washington, DC: Author, 1974.

Anastasi, A. (1976) *Psychological testing* (4th ed). New York: Macmillan.

Bakan, D. (1966) The test of significance in psychological research. *Psychological Bulletin*, 66, 423-437.

Bauernfeind, R.H. (1968) The need for replication in educational research. *Phi Delta Kappan*, 50, 126-128.

Bereiter, C. (1965) Issues and dilemmas in developing training programs for educational researchers. In E. Guba & S. Elam (Eds) *The training and nurture of educational researchers*. Bloomington, Ind: Phi Delta Kappa.

Bolles, R.C. (1962) The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports*, 11, 639-645.

Campbell, D.T. & Stanley, J.C. (1966) *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Carver, R.P. (1974) Two dimensions of tests: Psychometric and edumetric. *American Psychologist*, 29, 512-518.

Carver, R.P. (1976) Letter to the Editor. *Educational Psychologist*, 12, 96- 97.

Carver, R.P. (in press) Sense and nonsense about generalizing to a language population. *Journal of Reading Behavior*.

Chandler, R.E. (1957) The statistical concepts of confidence and significance. *Psychological Bulletin*, 54, 429-430.
Chase, L.J. & Tucker, R.K. (1976) Statistical power: Derivation, development, and data-analytic implications. *Psychological Record*, 26, 473-486.
Clark, C.A. (1963) Hypothesis testing in relation to statistical methodology. *Review of Educational Research*, 33, 455-473.
Coats, W. (June 1970) A case against the normal use of inferential statistical models in educational research. *Educational Researcher*, 6-7.
Cohen, J. (1977) *Statistical power analysis for the behavioral sciences* (Rev ed). New York: Academic Press.
Coleman, E.B. (1964) Generalizing to a language population. *Psychological Reports*, 14, 219-226.
Cronbach, L.J. (1975) Beyond the two disciplines of scientific psychology. *American Psychologist*, 30, 116-127.
Cronbach, L.J. & Snow, R.E. (1977) *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington.
Gold, D. (1969) Statistical tests and substantive significance. *The American Sociologist*, 4, 42-46.
Hays, W.L. (1963) *Statistics*. New York: Holt, Rinehart & Winston.
Hebb, D.O. (1966) *A textbook of psychology*. Philadelphia: Saunders.
Katzer, J. & Sodt, J. (1973) An analysis of the use of statistical testing in communication research. *Journal of Communication*, 23, 251-265.
Kish, L. (1959) Some statistical problems in research design. *American Sociological Review*, 24, 328-338.
Lykken, D.T. (1968) Statistical significance in psychological research. *Psychological Bulletin*, 70, 151-159.
Melton, A.W. (1962) Editorial. *Journal of Experimental Psychology*, 64, 553-557.
Morrison, D.F. & Henkel, R.E. (1969) Significance tests reconsidered. *The American Sociologist*, 4, 131-140.
Morrison, D.F. & Henkel, R.E. (1970a) Significance tests in behavioral research: Skeptical conclusions and beyond. In D.E. Morrison & R.E. Henkel (Eds) *The significance test controversy – A reader*. Chicago: Aldine.
Morrison, D.F. & Henkel, R.E. (1970b) *The significance test controversy – A reader*. Chicago: Aldine.
Nunnally, J.C. (1960) The place of statistics in psychology. *Educational and Psychological Measurement*, 20, 641-650.
Nunnally, J.C. (1975) *Introduction to statistics for psychology and education*. New York: McGraw-Hill.
Rozeboom, W.W. (1960) The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 57, 416-428.
Shulman, L.S. (1970) Reconstruction of educational research. *Review of Educational Research*, 40, 371-393.
Sidman, M. (1960) *Tactics of scientific research: Evaluating experimental data in psychology*. New York: Basic Books.
Smith, N.C. (1970) Replication studies: A neglected aspect of psychological research. *American Psychologist*, 25, 970-975.
Stevens, S.S. (1968) Measurement, statistics, and the schemapiric view. *Science*, 161, 849-856.
Stevens, S.S. (1971) Issues in psychophysical measurement. *Psychological Review*, 78, 426-450.

Tukey, J.W. (1962) The future of data analysis. *Annals of Mathematical Statistics*, 33, 1-67.
Tyler, R.W. (1931) What is statistical significance? *Educational Research Bulletin*, 10, 115-118; 142.
Wilson, K.V. (1961) Subjectivist statistics for the current crisis. *Contemporary Psychology*, 6, 229-231.
Winch, R.F. & Campbell, D.T. (1969) Proof? No. Evidence? Yes. The significance of tests of significance. *The American Sociologist*, 4, 140-143.